Applying FDR control subsequently to large scale one-way ANOVA testing in proteomics: practical considerations

Thomas Burger
Univ. Grenoble Alpes, CNRS, CEA, INSERM, ProFI, EDyP, F-38000 Grenoble, France
https://orcid.org/0000-0003-3539-3564; thomas.burger@cea.fr

Abstract

In discovery proteomics, as well as many other "omic" approaches, more than two biological conditions or group treatments can be compared in the one-way Analysis of Variance (OW-ANOVA) framework. The subsequent possibility to test for the differential abundance of hundreds (or thousands) of features simultaneously is appealing, despite requiring specific statistical safeguards, among which controlling for the False Discovery Rate (FDR) has become standard. However, using an FDR control procedure in a context where other multiple test corrections may also be involved because of the number of compared groups raises practical difficulties. This article surveys means to circumvent them, thanks to various practical data processing scenarios.

Keywords: Quantitative proteomics; Data processing; Biomarker discovery; One way Analysis of Variance (OW-ANOVA); False Discovery Rate (FDR); Post-Hoc Tests (PHTs).

1. Introduction

In its most general form, the biomarker selection problem amounts to select some features that can be significantly associated to a difference of biological status, with the objective to subsequently refine decision making (diagnostic, prognostic, patient follow up, etc.). In practice, one of the most widespread and efficient way to achieve biomarker selection is to perform a series of statistical tests (one for each feature available) and to retain those few that pass a user-defined significance threshold. This conceptual simplicity makes the procedure easily automatable, even though, in practice, two obstacles may raise questions and hinder the interpretation of the resulting biomarkers.

The first obstacle derives from the increase of instrumental throughputs: Many omics technologies provide quantitative measurements about hundreds (or even thousands) of features for each sample analyzed. From a biological perspective, this near exhaustiveness is a bliss, however, it also increases the chances that, for some of the putative biomarkers, random fluctuations concur with the change of biological status, hereby leading to artefactual significances; and thus to misleading biomarkers (a.k.a., false positives or false discoveries). To cope for this, it has become standard to control for the so-called False Discovery Rate (or FDR). FDR control procedures form a class of Multiple Test Corrections (MTC), which aims at adjusting the significance value used to retain putative biomarkers as a function of the total number of measured features. Therefore, in the end, the increment of omics measurement does not lead to an increment of false discoveries. Unfortunately, the theory underlying FDR control is not trivial and its correct application to biomarker selection remains at the center of an increasing body of literature.

The second obstacle derives from the increasing complexity of the biological questions, and more specifically, from the increasing complexity of the associated experimental designs. It its simplest form, one seeks for biomarkers discriminating two biological conditions, *e.g.*, Healthy vs. Disease (in a clinical context), or Wild-type *vs.* Mutant (in a fundamental research context). However, more refined experimental designs are increasingly popular: Multiple treatments *vs.* a reference condition, time-course (that should not be confused with longitudinal analyses, for which specific temporal models are used [1]), strata (*e.g.*, samples belonging to patients with different degrees of illness, etc.), or multi-

factorial designs (involving for instance a combination of severity, treatment and responsiveness). All these experimental designs share the need to compare $K > 2$ biological conditions. While an important body of applied statistic literature is dedicated to this subject, the most well-known statistical framework to address such experimental designs is that of the Analysis Of Variance (or ANOVA). The reason for the popularity of ANOVA is twofold [2][3]: first, it is a general enough framework to fit a variety of experimental designs; second, it generalizes the well-known t-test in an intuitive way (among others, notably when used in the fixed effect and OW setting), making its interpretation easier. However, beyond this superficial simplicity, some technical difficulties appear, notably with the downstream application of so-called Post-Hoc Tests (PHTs) that classically accompany fixed-effect OW-ANOVA.

If one depicts a dataset as a table where the rows correspond to the features (the putative biomarkers) and the columns correspond to the samples; then our first obstacle relates to the increment of the number of rows, while the second obstacle relates to the increasingly complex structure that binds the columns together. As previously mentioned, many solutions were proposed to each obstacle taken independently, among which FDR control (for the first obstacle) and ANOVA (for the second one) are the preferred ones. Unfortunately, the combined application of FDR control and ANOVA is intricate as PHTs include a form of MTC that relates to the number of compared biological conditions. Broadly speaking, FDR control amounts to a column-wise MTC while a PHT amounts to a row-wise one. It is thus essential to orchestrate them in a manner that remains statistically rigorous and biologically relevant. This article reviews the various ways to do so in a proteomics context (yet, its conclusions can easily be transposed to other omics settings).

2. Background

2.1. On FDR

Because of the important inter-individual variance of the proteome, but also because of the classically small number of available assays, any shortlist of putative biomarkers is expected to contain at least few false positives (*i.e.*, proteins for which the differential abundance experimentally observed does not convey any biological meaning). Their precise proportion (*i.e.*, their precise number divided by the shortlist length), classically referred to as the FDP (False Discovery Proportion), is unfortunately not accessible. However, since Benjamini and Hochberg seminal work [4], we have known that the mathematical expectation of the FDP (termed the False Discovery Rate or FDR) can be conservatively estimated; or controlled at a given level of risk α, using adjusted p-values [5] (or equivalently, q-values [6]).

In practice, the risk level α is often conflated with the FDR itself, as in sentences like "Differentially abundant proteins were validated at 1%FDR". In addition to advocate for a vocabulary simplification [7], [8] with respect to biostatistics theory [9], this conflation tends to reduce FDR framework to a feature selection method embedding its own multiple test correction. As such, the FDR control procedure has progressively stepped out of its original and prime role, namely assessing the statistical significance of a set of selected features, and has gained two other roles.

The first additional role is that of a quality control measure: The tuning to a default value (typically 1% in label-free discovery analyses) is more justified by the standardization capacities it yields (hereby providing a ground level to compare different analyses from different labs) than by advanced statistical reasons. The second additional role is that of practical filter: It helps the researcher focusing on an acceptable number of proteins by filtering out the others.

Finally, FDR control is now motivated by three roles: Statistical Rigor (SR), Quality Control (QC) and Practical Filter (PF). Ideally, these three motivations concur: By filtering the list of putative biomarkers to 1%, hereby fitting with the QC standards, one brings statistical support to those deemed differentially abundant while substantially reducing their number to an order of magnitude that is compatible with the constraints of post-proteomic validation wet-lab experiments. However, every

practitioner has ever faced a situation where this classical FDR threshold leads to a list of biomarkers, which is either far too short or far too long with respect to the anticipated validation experiments. Of course, it is possible to adjust the list by different tunings on preliminary filters (modifying the minimal number of peptides per protein, adjusting the minimal acceptable fold-change[1], etc.), but if it is insufficient, one may be tempted to adjust the FDR threshold to unrealistically low or high thresholds. This illustrates well that in practice, the tuning of the cutoff often results from a negotiation between the QC and the PF roles, while the SR one is simply assumed inherent to the methodology. Although not absurd in a simple experimental design like binary design (*e.g.*, Healthy vs. Disease) this view reaches its limit with more complex ones, which calls for a more explicit prioritization of the above mentioned roles.

2.2. On fixed-effect OW-ANOVA

Essentially, fixed-effect OW-ANOVA is a generalization of the t-test. Based on similar parametric assumptions (notably, the measurements should follow a normal distribution with equal variance in each group), one tests whether the means of an arbitrary number of groups equates, or not (while classical t-test can only handle two groups). More formally, while, the t-test null hypothesis reads:

$$\mathrm{H}_0^{two-means}: \mu_1 = \mu_2,$$

where μ_i is the mean of group i, the OW-ANOVA null hypothesis reads:

$$\mathrm{H}_0^{all-means}: \mu_1 = \mu_2 = \cdots = \mu_K,$$

where K is the number of groups. As with any statistical test, what we are really interested in is the rejection of $\mathrm{H}_0^{all-means}$, that is, cases where the K groups can reasonably be suspected to have some of their means different from the others. For those cases, textbooks in statistics suggest to apply post-hoc tests (PHTs) as to find which groups cause this difference.

Essentially, a PHT is a tool that combines two distinct statistical steps: First, a series of other tests which aim at refining $\mathrm{H}_0^{all-means}$; for instance, Tukey-Kramer HSD (Honest Significance of Difference) and Scheffé's PHTs propose to test $\mu_i = \mu_j, \forall i, j \leq K$, to find which pairs of groups display significant differences; similarly, Dunnett's proposes to test $\mu_1 = \mu_j, \forall\, 1 < j \leq K$, to check which groups are significantly different from the reference group 1 (see [14] for a thorough description of the available PHTs). Second, a multiple test correction (MTC) is applied, to avoid that the inflation of K leads to an inflation of significant tests. For instance, if one compares a reference treatment to 100 alternative treatments that are in fact placebos, we can expect one of them to be significant by chance at a significance level of 1%. To cope for such effect, many PHTs encompass a Family-Wise Error Rate correction (FWER, [15]), that is a correction broadly akin to that of Bonferroni's, despite subtle variations from one to another.

However, the OW-ANOVA+PHT workflow has two flaws that are usually not described in textbooks. First, significant rejection of $\mathrm{H}_0^{all-means}$ may not be confirmed by the downstream PHT and conversely, a significant rejection of a couple of means equality can be observed despite the non-rejection of $\mathrm{H}_0^{all-means}$ in the first place [16], which calls for alternative workflows [14]. Second, the conflation of the test series (those aiming at refining the omnibus test of $\mathrm{H}_0^{all-means}$) and the MTC into a single tool encourages a form of narrow-mindedness [17]. For instance, what if one is only interested in the pairwise comparisons of the form $\mu_i = \mu_{i+1}$? If one seeks proteins which are differentially abundant

[1] Applying a cut-off on the fold-change is a common practice in proteomics, however, if improperly achieved, it can hinder the statistical rigor. Notably, it has been demonstrated that applying such a cut-off downstream of the FDR is invalid [10]. It remains possible to incorporate the fold-change cut-off in the statistical test [11], but the resulting p-value interpretation can be a bit more difficult. Moreover, depending on how it is conducted, it can even lead to involuntary p-value hacking [12], so that applying such a filter prior to differential abundance testing constitute a nice trade-off [13].

within all these pairs (without exception), one should not correct for multiple testing using a FWER. Conversely, if from the beginning, one is only interested in comparing the new treatments or placebos against the reference treatment, why should $\mathrm{H}_0^{all-means}$ be tested in the first place? It would be equally sounded to consider as many pairwise t-tests as necessary, and then to apply a well-chosen MCT. Naturally, the writers of most of the textbooks were aware of this, but we should keep in mind that their first editions have often been anterior to the big data era, or even to the desktop computer. In those times, most of the computations were performed manually, or, at best, using a pocket calculator. In such circumstances, proposing a straightforward workflow that minimized the computations (*e.g.*, "spend time to deal with the pairwise comparisons only for the cases where an omnibus test has allowed you to reject $\mathrm{H}_0^{all-means}$") was probably more important than to account for the great diversity of experimental designs allowed by omics technologies (which were still to be created).

2.3. On constrast-based testing at scale

In omics biology, Limma [18] is probably the most popular implementation of ANOVA. The reason essentially lies in its moderated statistics, which allows for a better estimation of the variance of the putative biomarkers in a context where biological samples may be scarce. However, the influence of Limma goes beyond this.

First, it popularized the ANOVA serialization, *i.e.*, the testing of hundreds (or even thousands) of omics features (historically microarray probes) in an independent way. This approach has now become so mainstream that the warnings in Limma documentation are classically ignored. More precisely, according to the authors [19] combining FDR control (as classically required when testing simultaneously so many putative biomarkers) and FWER based MTC resulting from the combinatoric of pairwise comparisons is an issue. To cope for it, they have proposed four options, all of which with restrictions on the condition of applications or with lack of mathematical support (see below for a more refined discussion).

Second, Limma has made the power of ANOVA more explicit as a general framework to decompose the variance of a population into several groups. Although obvious to scholar statisticians, practitioners tend to conflate the mathematical framework of ANOVA, and the most widespread statistical test it provides (the omnibus F-test of mean equality, a.k.a. all-means F-test, a.k.a. the OW-ANOVA test). However, many null hypotheses other than $\mathrm{H}_0^{all-means}$ can be tested in the ANOVA framework. It is possible to test the equality of the groups' variances, to test the data fit to a regression model involving random or mixed effects [20], *etc.*: as long as the gaussianity and the homoscadisticity assumptions are respected together with the independence of the decomposition residues, it is possible to decompose the variance of the dataset so as to subsequently construct a test. Pushing the ease-of-use logic to its maximum in an applicative context, Limma authors has provided routines to automatically decompose the variance according to the biological questions at stake, allowing for an intuitive manipulation of the so-called "contrasts" by the practitioners.

Contrasts are of particular interest to unveil in an intuitive manner the mechanics underlying the classical fixed-effect OW-ANOVA workflow (all-means F-test + PHTs). More precisely, it is highlighting to note that the all-means F-test on K groups can be recovered from $K-1$ pairwise contrasts (that must nonetheless be chosen to respect a so-called orthogonality property [16]). Therefore, in a modern context of powerful computational resources, why bothering with a preliminary omnibus test, if from the very beginning, the practitioners know their practical interest will subsequently focus on specific contrasts?

Finally, tools like Limma advocate for a rational separation between the various constituents of the ANOVA workflow: the contrasts of interest; the row-wise or protein-wise MTC, which requirement results from the multiplicity of contrasts tested; and the column-wise or contrast-wise MTC, which requirement results from the number of omic features that are concomitantly but independently

tested. Notably, in a context where the joint application of both MTCs is still a research subject in theoretical biostatistics, this clear separation is instrumental to exhibit various scenarios, which provide a trade-off between statistical guarantees and applicability in function of the underlying biological question. However, it should be noted that these considerations are independent of the moderated statistics features for which Limma is most known. In other words, whether this moderated variance is used to increase the statistical power or not has no influence on how the practitioners should care about the interferences between PHTs and FDR. Likewise, it should not influence practitioners in their choice of the most adapted scenario among those hereafter described.

3. Various scenarios

Hereafter, the following notations are used: N refers to the number of proteins being tested; n (respectively, n_q) refers to the number of proteins being selected, globally (respectively, in contrast q); Q denotes the number of contrasts (which should not be confused with K, the number of groups or of biological conditions, see Figure 1a).

3.1. All-means F-tests + FDR + PHTs

A first intuitive scenario is to apply N omnibus tests on the N proteins, to select the n ones with the smallest p-values, with n so that the FDR is controlled at a user-defined threshold, and then, to apply the PHTs on the original dataset, yet restricted to these n proteins only (see Figure 1b).

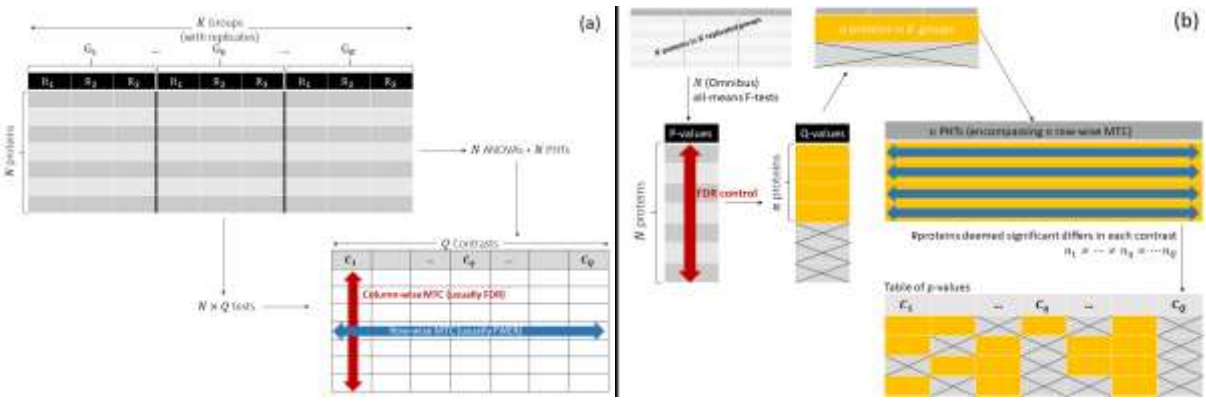

Figure 1: (a) schematic representations of the data and contrast tables with a recap of the mathematical notations. (b) All-means F-test + FDR + PHTs

With this scenario, it is important to understand that subsequently to the PHTs, the number of differentially abundant proteins may largely differ from one contrast to another, which may compromise the PF role of the feature selection process. Conversely, some proteins may display differential abundance for zero, one or even several contrast(s). Therefore, at the end of the process, the number of discovered differential abundances is strictly speaking not FDR-controlled anymore. Thus, the SR role of FDR is somehow lost, in spite of a selection process which stringency pleas for a certain conservatism. In this context, it thus remains important from a statistical viewpoint to tune the FDR threshold accurately. In fact, if n is too large, the number of retained differential abundances (among the n proteins and Q contrasts) should require an additional MTC, despite the preliminary FDR-based filter. Similarly, the more different the resulting list of biomarkers of interest is from the intermediate one (after FDR control but before PHTs), the more the SR of the whole workflow is compromised. As a result, with this data processing scenario, it is essential to tune the FDR threshold so that n is small. In other words, depending on the application, it is not possible to stick to the classical 1%FDR, which is so important in the QC role.

To summarize, with this scenario, the FDR thresholding may lose both its PF and QC roles while only approximately fulfil its SR role. Considering, it should not be promoted for most of the application cases. However, as an easy-to-reproduce workflow, it can be of interest when a proteomic experiment

is conducted at a very exploratory stage, to find if some proteins are changers between various groups, so that a up/down regulation mechanism can be hypothesized. However, it is important to notice that this workflow is more designed to reduce the number of hypotheses to subsequently explore (originally, as many as $n \times Q$), rather than to exhibit biomarkers with a statistical significance.

3.2. All-means F-tests +PHTs+FDRs

An alternative to the previous scenario is to swap PHTs and FDR control. As described on the path A illustrated on Figure 2, the N omnibus tests are followed by as many PHTs, leading to a table of $Q \times N$ p-values, and Q FDR controls are finally applied to provide a subset of Q lists of respectively $n_1, n_2, \dots, n_Q$ proteins. As with the previous scenario, the values $n_1, n_2, \dots, n_Q$ may not be equal, and even in case of equality, the Q lists of differentially abundant proteins may not be the same from one contrast to another. However, and in contradiction with the previous scenario, the practitioner can tune the Q FDR thresholds to have broadly equal $n_1, n_2, \dots, n_Q$. Although questioning from a SR viewpoint, doing so may be useful in a PF aim. Nevertheless, if the FDR thresholds are all tuned to a same value dictated by QC standards, then the scenario is compatible with both the SR and QC roles (briefly, as no processing occurs downstream of the FDR control, it remains valid).

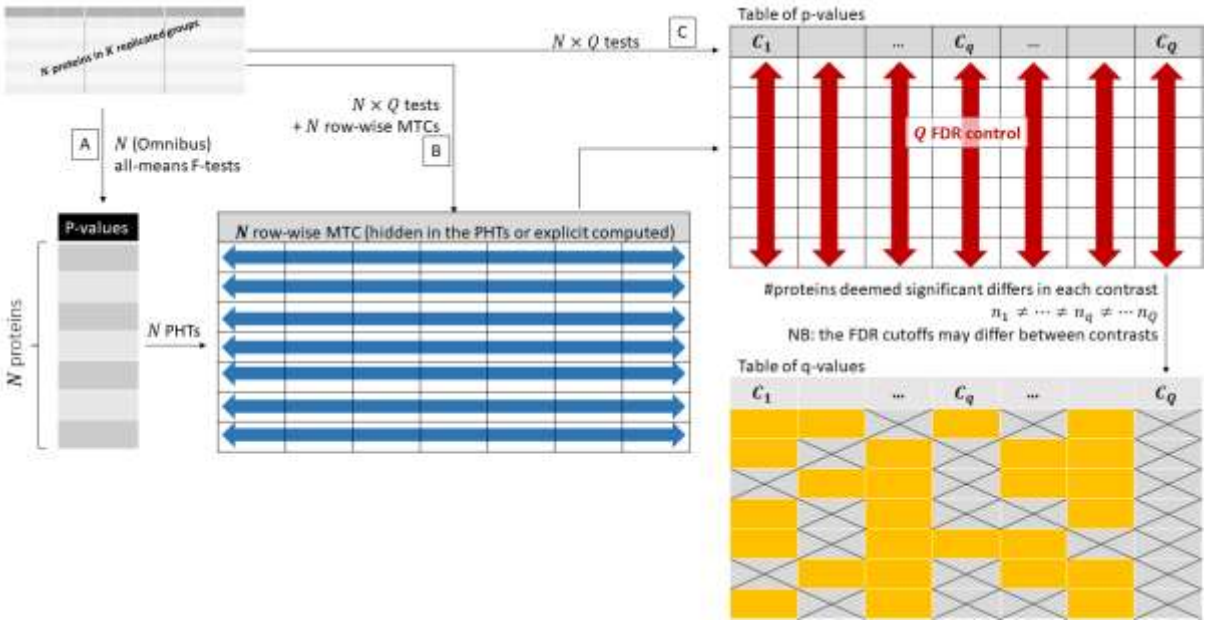

Figure 2: [A] All-means F-test + PHTs+ FDRs; [B] Multiple Contrast tests + FWERs + FDRs; [C] Multiple Contrast tests + FDRs (referred to as the "separated" option in Limma)

As such, this second scenario can be seen as the ultimate one, in the sense that it is sufficiently flexible to manage a trade-off between the three roles of FDR. However, studying this scenario at the light of contrast-based analysis raises questions. As mentioned earlier, if an omnibus test is systematically followed by PHTs, it probably means that in fact, the omnibus test is not needed, making a workflow like "Multiple Contrast tests + protein-wise FWER controls + contrast-wise FDR controls" probably more sensible, as illustrated on the path B of Figure 2. For sake of clarity, let us rely on Bonferroni correction to impersonate this FWER control (even though this choice may not lead to an exactly equivalent scenario, mainly because of the various ways to correct the FWER across the PHTs, the difference is not sufficiently important to be a matter of concern here).

Concretely, applying Bonferroni correction amounts to replace each of the $n \times Q$ p-values (noted p_i) resulting from the testing of the Q contrasts by $p_i' := \min(Q \times p_i, 1)$. In other words, each p-value is multiplied by Q (and if necessary thresholded to 1), which leads to a dramatic reduction of the number of possible differential abundances (in other words, this scenario is overly conservative, notably when the number of contrasts increase). Of course, this is the main reason why the approach is in line with

the SR role, but at some point, it makes sense to question its interest. Obviously, when considering the influence of K treatments on a specific biomarker, a Bonferroni-like correction is mandatory to avoid that the risk of selecting a placebo by chance increases with K. However, when considering a time-course experiment, if one increases the number of time stamps by a twofold (*e.g.*, by performing a biopsy every 2 hours instead of every 4h), we do not expect to observe a double number of proteins which global profile appears interesting. Conversely, if a globally differentially abundant protein displays two times more differential contrasts, it will not fundamentally alter its potential for subsequent investigations. As a result, depending on the biological question, this intermediate protein-wise FWER control may be either necessary, or overly and unnecessarily conservative. As dealing with overconservative statistical procedures is a recurrent problem for practitioners, the rest of this article describes various scenarios, where the specificities of the experimental design car be leveraged to by-pass the intermediate row-wise MTCs (or at least, to make less conservative).

3.3. Separated contrasts+FDRs

Boldly skipping the row-wise MTCs in the previous scenario (as illustrated on the path C of Figure 2) may appear incompatible with statistical guidelines. However, it is not necessarily the case if one avoids subsequent over-interpretation. In fact, this scenario is rather common as it is implemented in Limma (under the method "`separate`" in the `decideTests` function) where it is the default method. As explained in the documentation, this method "[…] *does multiple testing for each contrast separately* […]. *Using this method, testing a set of contrasts together will give the same results as when each contrast is tested on its own. The great advantage of this method is that it gives the same results regardless of which set of contrasts are tested together. The disadvantage of this method is that it does not do any multiple testing adjustment between contrasts. Another disadvantage is that the raw p-value cutoff corresponding to significance can be very different for different contrasts* […] [as with "All-means F-test + PHT+ FDRs" in Section 3.2]. *This method is recommended when different contrasts are being analysed to answer more or less independent questions.*"

Using this method, the most difficult point is to determine whether the questions at stake are sufficiently independent. Notably, comparisons involving independent biological conditions do not imply that the associated contrats (or questions) are independent. As a counter-example, just think about the testing of multiple treatments: although chemically independent, the manifold of treatments requires MTC. Finally, with this scenario, the SR role of FDR control may be preserved as long as one considers sufficiently independent questions. As for the PF and QC roles, this scenario stands exactly at the same point as that of Section 3.2.

3.4. All-means F-test + FDR+ Clustering

Let us now depict another situation where it is possible to solve the entanglement of row-wise and column-wise MTCs by getting rid of the less necessary ones (*i.e.*, the row-wise MTCs, as when testing simultaneously hundreds of proteins, it is simply impossible to bypass the FDR control). This scenario is well suited to answer a "global" question, *i.e.*, a question that does not translate into a single specific contrast. Notably, when studying time-courses with so many time-stamps that exhaustive pairwise comparisons become intractable, one should mainly look at the global trend of each protein. In other words, one essentially wants to discriminate between proteins whose abundance remains stable despite random fluctuations, and those whose abundance changes significantly, in global terms. Of course, one may be also interested in whether each variation of abundance amounts to an up or down-regulation; as well as whether it occurs at the beginning, at the middle or at the end of the time-course. However, depending on the application, these refinements may remain of qualitative nature (*i.e.*, devoid of associated significance values), without flawing the biological reasoning and results.

In those cases, an omnibus test like the all-means F-test is sufficient to provide a global significance value to each protein profile, while complementary information can be emphasized with exploratory and descriptive tools, like cluster analysis. As an illustration, Figure3a schematizes several clusters obtained on averaged and standardized protein profiles (*i.e.*, each protein profile is constructed by

averaging all the samples' value within each group; then the profile is centered on 0 and scaled to a unitary variance [21]). By combining these clusters with the result of the all-means F-test, it is possible to spot proteins which profiles (whatever their shape) are significantly different from a flat profile according to the user-defined FDR threshold.

For most proteomics experiments that will need post-analytics validations, results of this type are sufficient. However, it is important to avoid over-interpretation when describing them. For instance, if among a cluster of proteins exhibiting an early downfall in their profile, one is significantly differentially abundant according to the omnibus test, it is not possible to conclude that the early down-regulation itself is significant. Saying so would amount to transfer the significance associated to the global rejection of $H_0^{all-means}$ to a more specific question (namely, the early down-regulation), which would constitute a statistically unsupported claim. However, it remains nonetheless possible to conduct downstream experiments to assess further this early down-regulation. Therefore, as long as one discriminates between conclusions that require statistical significance supports, and intermediate evidence along the course of a biological investigation (which may not require a similar support as experimental confirmations are expected), it is possible to keep a safe and rigorous line of reasoning. Most importantly, doing so makes it possible to reconcile the three roles of FDR control: SR (as long as one does not over-interpret and accepts an answer to a broader question), PF and QC (notably, managing a single selection process makes it easier to reach an agreement between the last two roles, as with a single binary contrast).

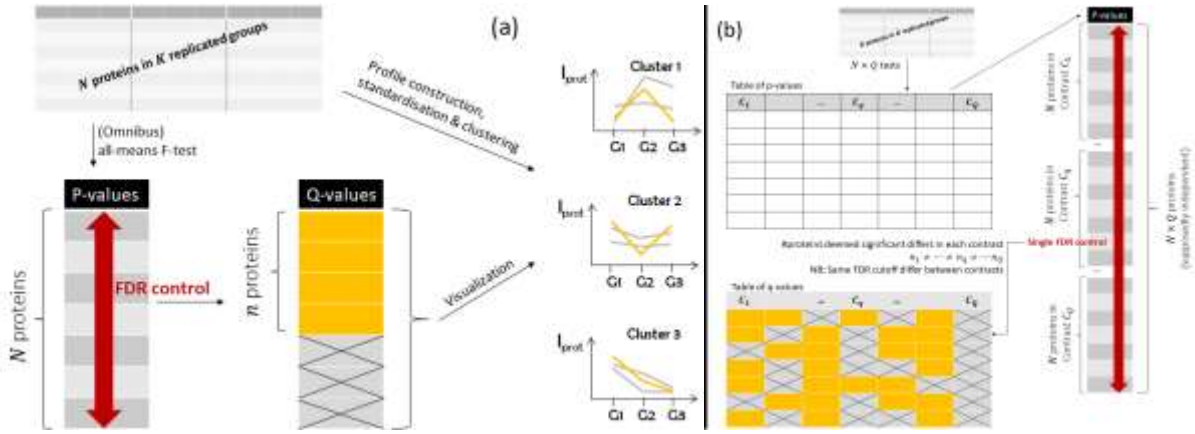

Figure 3: (a) All-means F-test + FDR + Clustering; (b) Single FDR control over stacked contrasts (Global option Limma)

3.5. Stacked Contrasts + single FDR control

In the cases which do not fall in the previous categories (*i.e.*, the questions are not sufficiently independent and they require going beyond a global approach), the simplest workaround is to stack all the contrasts in a single large contrast, so that the cumbersome combinations of row-wise and column-wise MTCs is replaced by a single column-wise MTC (see Figure 3b). In other words, a given protein yielding two p-values for two different contrasts are considered exactly as two p-values resulting from two proteins tested in a single contrast. This strategy has also been promoted by Limma package, through the method referred to as "`global`" in the `decideTests` function. This approach has many advantages, which are listed in Limma user's guide [19]. Their paraphrasing here would be of little interest, but to summarize, the PF and QC roles classically endowed to FDR control procedure are fully accounted for.

Unfortunately, depending on the column-wise MTC, this approach may lack theoretical supported. Notably, when this MTC amounts to an FDR control (which is the most classical approach in proteomics), the authors say the following: "[…] *there is no theorem which proves that* `adjust.method="BH"` *in combination with* `method="global"` *will correctly control the false*

discovery rate for combinations of negatively correlated contrasts, however simulations, experience and some theory suggest that the method is safe in practice." Although the absence of theoretical guarantee should not systematically discard the practical application of an otherwise interesting approach, relying on qualitative and simulation-based empirical arguments only is not sufficient to trust in the SR role fulfillment. A middle of the road approach is to stack the Q p-value vectors manually and then to run a tool like CP4P [22]. CP4P makes it possible to check the conditions of application of BH FDR control through the visualization of the p-value calibration, as well as to some extent, to correct for minor calibration defaults, as detailed in [13]. Then, depending on the calibration assessment, the practitioner may decide whether to use this scenario (*i.e.*, a single FDR control on stacked contrasts, *i.e.*, `adjust.method="BH"` in combination with `method="global"` in Limma).

3.6. p-value aggregation across contrasts + single FDR control

In cases where the previous scenarios are not satisfactory, it is possible to rely on more refine approaches, which all share a common principle: Firstly, for each protein, aggregating all the p-values resulting from the Q contrasts into a single p-value; secondly, applying an FDR control on this single "aggregated contrast" (see Figure 4). Theoretically speaking, aggregating the multiple p-values resulting from multiple contrasts into a single protein-wise p-value can be seen as a form of MTC. Thus, by leveraging the appropriate procedure, one should recover the statistical rigor that may be missing with the above scenarios.

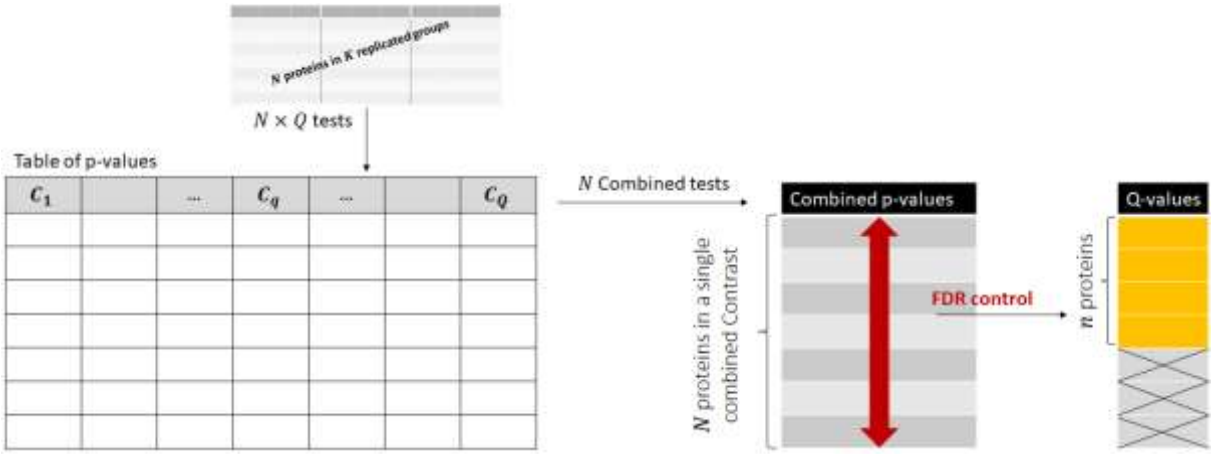

Figure 4: In this scenario, the various contrast p-values are aggregated for each protein so as to provide a single p-value for each, then FDR control is applied.

Many p-value combination methods endowed with proof of correct calibration exist in the literature (Harmonic mean [23], Fisher combination [24], etc.), but few of them can be interpreted as contrast summarization. However, among them, Simes procedure [15] is already used, through the so-called `hierarchical` method of Limma's *decideTests* function. Unfortunately, it suffers from the same drawbacks as the previous scenario (where the contrasts are stacked): Although adapted to many column-wise MTCs based on FWER, the interaction with FDR control approaches is still an open subject. Therefore, the same advice as above may be applied: The use of this approach should remain pending to an upstream assessment of BH procedure conditions of application, notably through CP4P.

However, beyond Simes procedure, other p-value aggregation methods can be thought of. Following a previously described scenario, let us assume to be mainly interested by a global view of differential abundance, as for instance, in a time-course. Following this line, a protein can be interesting because it displays a consistent trend, even if this trend is sometimes significant and sometimes not (for instance, the protein is constantly up-regulated, but this is not significant between some consecutive time stamps, while it is between others). Although informative, this may be too restrictive. One may be tempted to focus on the smallest p-value between the Q contrasts for each protein, as a marker of

its "best fold change". However, taking the minimum out of Q p-values requires a subsequent Šidák correction [25], which as-a-matter-of-factly corresponds to an FWER control:

$$p := 1 - \left(1 - \min_{1 \leq i \leq Q} (p_i)\right)^Q$$

After computing this single p-value for each protein, it becomes easy to subsequently control the FDR, as only one "aggregated contrast" remains. Conversely, if one is interested in the "worst fold change" among Q contrasts, it is possible to consider the max p-value. Although very conservative, this can be instrumental when comparing the weakest link among cascading events. Then, it is possible to adapt Šidák correction to maintain the correct calibration, and to compute p-values as follows:

$$p := \left(\max_{1 \leq i \leq Q} (p_i)\right)^Q$$

The interest of this approach is that depending on the question at stake, it is possible to rely on the important body of literature about how to combine statistical tests to simplify the interferences between the MTCs resulting from the multiple contrasts and those resulting from the manifold of proteins tested.

4. Discussion and conclusions

This article surveys the various ways to orchestrates multiple comparisons in the fixed-effect One Way ANOVA framework (using either an omnibus test or several specific contrasts) and the subsequent post-hoc tests with the needs for false discovery rate control. The cases where post-hoc tests and false discovery rate control interfere are discussed and several practical scenario where the post-hoc tests can be easier by-passed or simplified are presented.

The fixed-effect One Way ANOVA typical workflow (OW-ANOVA) composed of an all-means F-test and subsequent Post-Hoc Tests (PHTs) is particularly popular to process omics data. The reasons are threefold: First, its genericity, which makes it suitable enough to a large variety of designs and of biological questions. Second, its ease of understanding and interpretation, as a straightforward extension to the t-test. Third, its considerable importance from a historical perspective.

However, its systematic use may not be adapted to modern omics data: First, despite being rather generic, it is usually possible to find an alternative contrast-based workflow that is more specific to the design or biological question considered. Second, its computational efficacy has become of limited (or null) interest according to the modern computational resources. Finally, its PHTs difficultly orchestrate with the need for FDR control. All these advocate for a rational use of this OW-ANOVA workflow, and for the growing use of alternative scenarios, such as those described above.

When doing so, practitioners will notice that the use cases where the ANOVA PHTs and FDR controls need to be used concomitantly are in fact scarce. In those situations, it is mandatory to wonder about which roles are expected from the FDR control, between "statistical rigor", "practical filter" and "quality control". Notably, it can be accepted that the first one is lost (while it should classically be the prime one), as long as either (i) statistical significance is not necessary and the reported conclusions are not overlooked; (ii) complementary filters are used to enforce the statistical significance. However, it is most of the time simpler to rigorously use an FDR control and to adapt the statistical test to limit the need for PHTs (or more generally for protein-wise MTCs), as advocated in this review.

Finally, from the proteomics researcher viewpoint, the most difficult aspect resides in a paradigm shift. The systematic application of the classical OW-ANOVA workflow was mainly a technical computed-aided manipulation, which fulfilling granted the passing of statistical guidelines. On the contrary, this review invites researchers to abandon the traditional binary vision about statistical tools (the mandatory/approved ones *vs.* the forbidden/incorrect ones) and to incorporate their educated use

into their global scientific thinking. While doing so would be beneficial to any omics biology researcher, it is particularly so for proteomics experts, as their data reflects the inherent complexity of the analytical workflow at use. The complex missingness patterns resulting from mass spectrometry based analyses [26], the reduced number of generally available assays, the limited protein coverage of tryptic peptides and conversely, the presence of shared peptides between proteins, etc. All these experimental issues impact the data so that they can deviate from the ideal cases such as reported in statistical textbooks or in this review, so that compromises may be necessary. To best do so, proteomic researchers need this paradigm shift. It will notably lead them to better formulate the biological question and improve the associated design, while defining the exact level and type of statistical significance that are necessary to pursue their investigations. Although this shift requires a more interdisciplinary vision, on the long run, it is the easiest way to choose among the various possible statistical scenarios, the one that best fits the data.

Acknowledgement and fundings

The author wish to thank Hélène Borges, Yohann Couté and Virginie Brun for fruitful discussions. 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), LIFE project (ANR-15-IDEX-02) and MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

Reference

[1] Hedeker, D., & Gibbons, R. D. (2006). Longitudinal data analysis. Wiley-Interscience.
[2] Tsuji, T., Shiozaki, A., Kohno, R., Yoshizato, K., & Shimohama, S. (2002). Proteomic profiling and neurodegeneration in Alzheimer's disease. Neurochemical research, 27(10), 1245-1253.
[3] Eckel-Passow, J. E., Oberg, A. L., Therneau, T. M., & Bergen III, H. R. (2009). An insight into high-resolution mass-spectrometry data. Biostatistics, 10(3), 481-500.
[4] 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), 57(1), 289-300.
[5] Murray, M. H., & Blume, J. D. (2021). FDRestimation: Flexible False Discovery Rate Computation in R. F1000Research, 10(441), 441.
[6] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. The Annals of Statistics, 31(6), 2013-2035.
[7] Burger, T. (2018). Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. Journal of proteome research, 17(1), 12-22.
[8] Etourneau, L., Varoquaux, N., & Burger, T. (2021). Unveiling the links between peptide identification and differential analysis FDR controls by means of a practical introduction to knockoff filters. bioRxiv.
[9] Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. Test, 12(1), 1-77.
[10] Ebrahimpoor, M., & Goeman, J. J. (2021). Inflated false discovery rate due to volcano plots: problem and solutions. Briefings in bioinformatics, 22(5), bbab053.
[11] McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics, 25(6), 765-771.
[12] 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.
[13] Wieczorek, S., Gianetto, Q. G., & Burger, T. (2019). Five simple yet essential steps to correctly estimate the rate of false differentially abundant proteins in mass spectrometry analyses. Journal of proteomics, 207, 103441.
[14] Ruxton, G. D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. Behavioral ecology, 19(3), 690-693

[15]Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. Statistics in medicine, 33(11), 1946-1978.
[16]Tian, C. H. E. N., Manfei, X. U., Justin, T. U., Hongyue, W. A. N. G., & Xiaohui, N. I. U. (2018). Relationship between Omnibus and Post-hoc Tests: An Investigation of performance of the F test in ANOVA. Shanghai archives of psychiatry, 30(1), 60.
[17]Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33(5), 587-606.
[18]Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.
[19]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. See online version (14 November 2021), section 13.3, page 63. https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf
[20]Mercier, C., Truntzer, C., Pecqueur, D., Gimeno, J. P., Belz, G., & Roy, P. (2009). Mixed-model of ANOVA for measurement reproducibility in proteomics. Journal of proteomics, 72(6), 974-981.
[21]Futschik, M. E., & Kumar, L. (2009). Introduction to Mfuzz package and its graphical user interface. Analysis, 1-13.
[22]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.
[23]Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences, 116(4), 1195-1200.
[24]Fisher, R. A. (1992). Statistical methods for research workers. In Breakthroughs in statistics (pp. 66-70). Springer, New York, NY.
[25]Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62(318), 626-633.
[26]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.