VIEWPOINT

Uses and misuses of the fudge factor in quantitative discovery proteomics

Quentin Giai Gianetto, Yohann Couté, Christophe Bruley, Thomas Burger*
BIG-BGE (Université Grenoble-Alpes, CNRS, CEA, INSERM), Grenoble, France.
*thomas.burger@cea.fr

Abstract: Selecting proteins with significant differential abundance is the cornerstone of many relative quantitative proteomics experiments. To do so, a trade-off between *p*-value thresholding and fold-change thresholding can be performed thanks to a specific parameter, named fudge factor, and classically noted s_0. We have observed that this fudge factor is routinely turned away from its original (and statistically valid) use, leading to important distortion in the distribution of *p*-values, jeopardizing the protein differential analysis; as well as the subsequent biological conclusion. In this article, we provide a comprehensive viewpoint on this issue, as well as some guidelines to circumvent it.

Keywords: discovery proteomics, statistical analysis, differential abundance.

In quantitative discovery proteomics, once peptides have been identified, their quantitation signal extracted and processed (logarithmic transformation, filtering, normalization, missing value imputation, protein aggregation [1][2]), differential analysis can be conducted. Basically, it amounts to perform null hypothesis significance testing on each protein so as to decide if it is differentially abundant (or not) between the compared conditions, depending on a user-defined threshold (a specific *p*-value or an expected false discovery rate [3]).

During this procedure, it is also common to filter all the proteins which fold-change (*FC*) is too small between the two conditions to compare; as small *FCs* are more likely to badly generalize to different biological samples, and then are unlikely to be validated by post-proteomics experiments; or as they are usually thought not to be biologically significant. As such, it is very common to apply the following procedure: (1) Proteins are represented on a volcano-plot, that is a scatterplot with *FCs* on the X-axis, and scores related to the differential analysis (generally –log10(p) where p is the *p*-value from the statistical test) on the Y-axis; (2) "Vertical" thresholds are applied on the graph, to filter out too small *FCs*; (3) "Horizontal" threshold is applied to select small *p*-values (see for instance Fig. 1a). A derivation of this procedure is to merge these two thresholds into smooth curves (such as illustrated on Fig. 1c). The resulting volcano-plot is much nicer to look at, and it intuitively reduces boundary effects, as the sharp corners of the regions with proteins of interest are soften. This is why it has rapidly widespread in the proteomics community ([4][5][6] are examples of works that explicitly rely on it, a far greater number could be cited). As explained in the protocol of [7], this smooth thresholding is concretely implemented by some s_0 parameter that, according to [6] is inherited from the SAM-test. The SAM-test

is a variation of the classical *t*-test where this additional s_0 parameter is named *fudge factor* [8]. The authors of [7] rightfully pinpoints the fact that different tuning of s_0 leads to different tuning of the thresholding curve.

In this article, we show this procedure may lead to biased differential analysis and to unsupported biological conclusions if s_0 is not rigorously tuned. First, we go back to the origin of SAM-test, and provide explanation on the origin of the fudge factor. Second, we explain why it was tempting to turn it away from its original use and to adapt it to filter out proteins with too low *FC*. Then, we illustrate its misuse on real biological data. Finally, we go back to the original procedure, that is the double *p*-value and *FC* thresholding, and we explain why it is the correct way to perform differential analysis.

Some explanations on the fudge factor: Basically, the goal of any Student-like *t*-test is to weight the *FC* by the inverse of the standard deviation (noted σ) of the abundances in each condition. For any protein i, its statistics reads:

$$t_i = \left(\mu_i^A - \mu_i^B\right)\frac{1}{\sigma_i} = \frac{FC_i}{\sigma_i}$$

with $FC_i = \mu_i^A - \mu_i^B$ where μ_i^A and μ_i^B are the averages of log-transformed intensities in compared conditions A and B. However, in practice, σ_i is unknown, and only its estimate, noted s_i, can be used:

$$\hat{t}_i = \frac{FC_i}{s_i}$$

In case of too few replicates, the *t*-test exhibits a major weakness: The difference between σ and s, even if rather small, may be too important to provide accurate *p*-values. To illustrate this, let us assume that there is such a small difference, noted δs_i:

$$\sigma_i = s_i + \delta s_i$$

Let us investigate the consequences of this small difference on the quality of the *t*-statistics: If in the above definition, σ were involved in the numerator, then, the *t*-statistics would be underestimated by a factor of $1 - \delta s_i$, (or overestimated, if $\delta s_i < 0$). As δs_i is small, it would not be an issue. Unfortunately, σ_i is involved in the denominator, so that the ratio of the real and estimated statistics reads:

$$\frac{\hat{t}_i}{t_i} = \frac{FC_i/s_i}{FC_i/\sigma_i} = \frac{1/s_i}{1/(s_i + \delta s_i)} = \frac{s_i + \delta s_i}{s_i} = 1 + \frac{\delta s_i}{s_i}$$

Finally, in real life, the small imprecision of the estimated standard deviation leads to an overestimation of the *t*-statistics by a factor of $1 + \delta s_i/s_i$. Naturally, if the proteomics experiment is poorly reproducible, so that the empirical variance on the measured abundance for protein i is high, δs_i is much smaller than s_i, so that $1 + \delta s_i/s_i$ is close to 1, and the approximation is hardly an issue. However, in case of a proteomic experiment of high quality, where the quantification is very reproducible across the replicates, s_i is likely to be small, and possibly, smaller than 1. In such a case, $1 + \delta s_i/s_i$ may significantly deviate from 1, leading to strong *t*-statistics overestimation. In the worst case (if the measurements are perfectly reproducible, or if by chances, the abundance measurements coincide) the empirical variance is nil, so that $\delta s_i/s_i$ is infinite, leading to dubious infinite *t*-statistics.

To summarize, if the abundances measured in the different replicates of a proteomic experiment are rather reproducible, it is possible to have a fair estimate of the variance, i.e. $\sigma_i \approx s_i$, while, on the contrary,

$$\frac{1}{\sigma_i} \neq \frac{1}{s_i}$$

In the end, this may lead to estimated *t*-statistics that strongly differ from the real one. This issue is well-known in numerical analysis, as it occurs whenever one divides by a small quantity which is fit with some

imprecision. This is why numerous computational tricks are used to face it. The most popular one is named Tikhonov regularization [9]; it is based on adding a small constant before computing the inversion.

To provide a solution to the instability of the *t*-test in omics sciences, where classically, standard-deviations are weakly estimated because of a small number of replicates, it has been proposed to apply a regularization trick: The authors of the SAM-test proposed to add a small constant s_0 to the estimated standard deviation. The idea is that, whatever the imprecision δs of the standard deviation estimate, it will be accounted for, as long as one has $\delta s_i \leq s_0$. In the worst case where for a specific protein i, all the measurements coincide, dubiously leading to $s_i = 0$, the *t*-statistics reads:

$$\tilde{t}_i = \frac{FC_i}{s_i + s_0} = \frac{FC_i}{s_0}$$

and does not reach an infinite value, as precisely justified in [8]. Beside the mathematical trick, let us note that s_0 has a nice practical interpretation: Whatever the measurements, the standard deviation cannot be considered smaller than s_0; thus, s_0 can be interpreted as an estimation of the volatility of the measurement.

Naturally, statisticians are well aware that tuning a parameter related to numerical stability may be difficult to practitioners. This is why, several works have thoroughly investigated the issue [10][11]. In addition, the authors of the SAM-test provided some guidelines, which have been translated into an automatic tuning procedure, available together with in the SAM-test R package (siggenes, function `fudge2()`, [12]). Let us remark that according to these guidelines, s_0 should stick to small values, and should be tuned only according to the very content of the dataset, regardless of any other "external" information. As a result, in the original publication, one has $s_0 = 3.3$ for a dataset where the mean intensities range between 0 and 25,000. However, on a proteomic dataset with log-transformed intensities, such as for instance those reported in [13], `fudge2()` advised $s_0 \approx 0.02$. This is significantly smaller than other values reported in proteomics literature (see for instance [4][5][6][14][15][16] where s_0 ranges between 0.4 and 4).

Consequences of deviating from the recommendations of [8]: Let us explore the results of tuning s_0 to an important value with respect to s, regardless of `fudge2()`. To make it more understandable, let us consider the extreme case, where s_0 would be so great that, in proportion, σ_i is immaterial, whatever the protein i:

$$\frac{s_i + s_0}{s_0} \approx \frac{\sigma_i + s_0}{s_0} \approx 1$$

As s_0 is the same for all the proteins of the dataset, the *t*-statistics reads:

$$\tilde{t}_i = \frac{FC_i}{s_i + s_0} \approx \frac{FC_i}{s_0}$$

for each protein i, and not only the few ones which by chance have a null empirical variance, such as described above. Thus, this new *t*-statistics provides the same ranking amongst the proteins that the *FC*; or more formally, $\forall$ couple of proteins (i,j), $\tilde{t}_i \geq \tilde{t}_j \Leftrightarrow FC_i \geq FC_j$. Finally, in this extreme situation, the *p*-value cut-off would be the same as if one uses the *FC*; rather than the *t*-statistics or the correctly tuned SAM-statistics. In other words, it is possible to make the *t*-statistics looks like the *FC* by tuning s_0 to some extremely great value with respect to s. Consequently, the use of s_0 has shifted from that of a mathematical tricks (Tikhonov regularization) the tuning of which could be related to the physics of the measurement process; to a parameter loosely tuned to account for a filtering on the *FC*.

Clearly, this is an extreme scenario. However, if s_0 is tuned with an in-between value, the behavior will be a mix between a regularizer accounting for the measurement variability, and a *FC* cutoff. This intuitively leads to the expected "smooth thresholds". However, and contrarily to what is suggested in [6], such use of the fudge factor is not supported by the original publication, so that the resulting *p*-value modifications are unsupported. On the contrary, it can be derived from [17], that an arbitrary tuning of s_0 can lead to spurious distortions of the values of the *t*-statistics, which provides solid arguments against an unsupported tuning.

Let us illustrate this on real data: We consider the dataset named *LFQRatio2* that can be found on the ProteomeXchange repository with the identifier PXD002370 (see suppl. material) and that accompanies the R package "cp4p" [13]. It is a series of six samples containing the same yeast lysate background in which the Sigma UPS1 equimolar mixture was spiked in, with different concentrations: Samples 4 to 6 received 2 times more UPS1 proteins than samples 1 to 3. On such a controlled dataset, all and only the UPS1 proteins are supposed to be differentially abundant. On its basis, it is possible to construct a volcano-plot, where the *p*-values are given by the Student *t*-test. The R code to download the dataset and display the volcano-plot is given in supplemental material. The horizontal and vertical thresholds are straightforward to draw, leading to the display of Fig. 1a.

Regarding the smooth curve threshold, things are more complicated. First, let us pinpoint that even if one uses the SAM-test, the Student *p*-values are still used, so that the volcano-plot does not change. However, one computes the SAM-test *p*-values and threshold them to discriminate differentially and non-differentially abundant proteins (this is equivalent to the horizontal threshold of Fig. 1a). Thus, the smooth curve only pictures this SAM-test-induced discrimination in the *t*-test/FC representation. Concretely, it can be demonstrated (see suppl. material) that if one tunes a confidence level $1-a$, then, the equation of the curves reads:

$$f(FC) = \begin{cases} -\log_{10}\left(2 \times \left[1 - F_{St_d}\left(t_a \times \left(1 + \dfrac{s_0}{\dfrac{FC}{t_a} - s_0}\right)\right)\right]\right) & \forall FC > t_a s_0 \\ -\log_{10}\left(2 \times \left[1 - F_{St_d}\left(t_a \times \left(1 + \dfrac{s_0}{\dfrac{FC}{-t_a} - s_0}\right)\right)\right]\right) & \forall FC < -t_a s_0 \end{cases}$$

where F_{St_d} is the Student distribution function, and where t_a relates to a:

$$a = \mathbf{P}\left(\frac{|FC|}{\sigma + s_0} \geq t_a | H_0\right)$$

(H_0 denotes the probability is computed for non-differentially abundant proteins);

The purpose is now to tune f to perform sensible thresholding. According to [6], s_0 "*sets a threshold for minimum fold change*", so that if one wants to mimic the thresholding of Fig. 1a, s_0 should be set to 0.5. Similarly, the thresholding on the *p*-values should simply relate to the confidence level $1-a$, so that it makes sense to tune it to 97.5%. The result is displayed on Fig. 1b. Surprisingly, it does not look as expected: In fact, contrarily to what is classically assumed, the *FC* cut-offs do not relate to $\pm s_0$, but to $\pm t_a s_0$ which are asymptotes for f. In a similar way, t_a is directly involved in the horizontal asymptote of f. This is really an issue as the practitioner cannot easily estimate t_a: this value relates to the SAM-statistics, while on the contrary, the user's confidence level relates to the *t*-statistics. Consequently, by

replacing the classical horizontal and vertical thresholds by smooth curves, the practitioner has replaced easy-to-interpret parameters by abstruse ones.

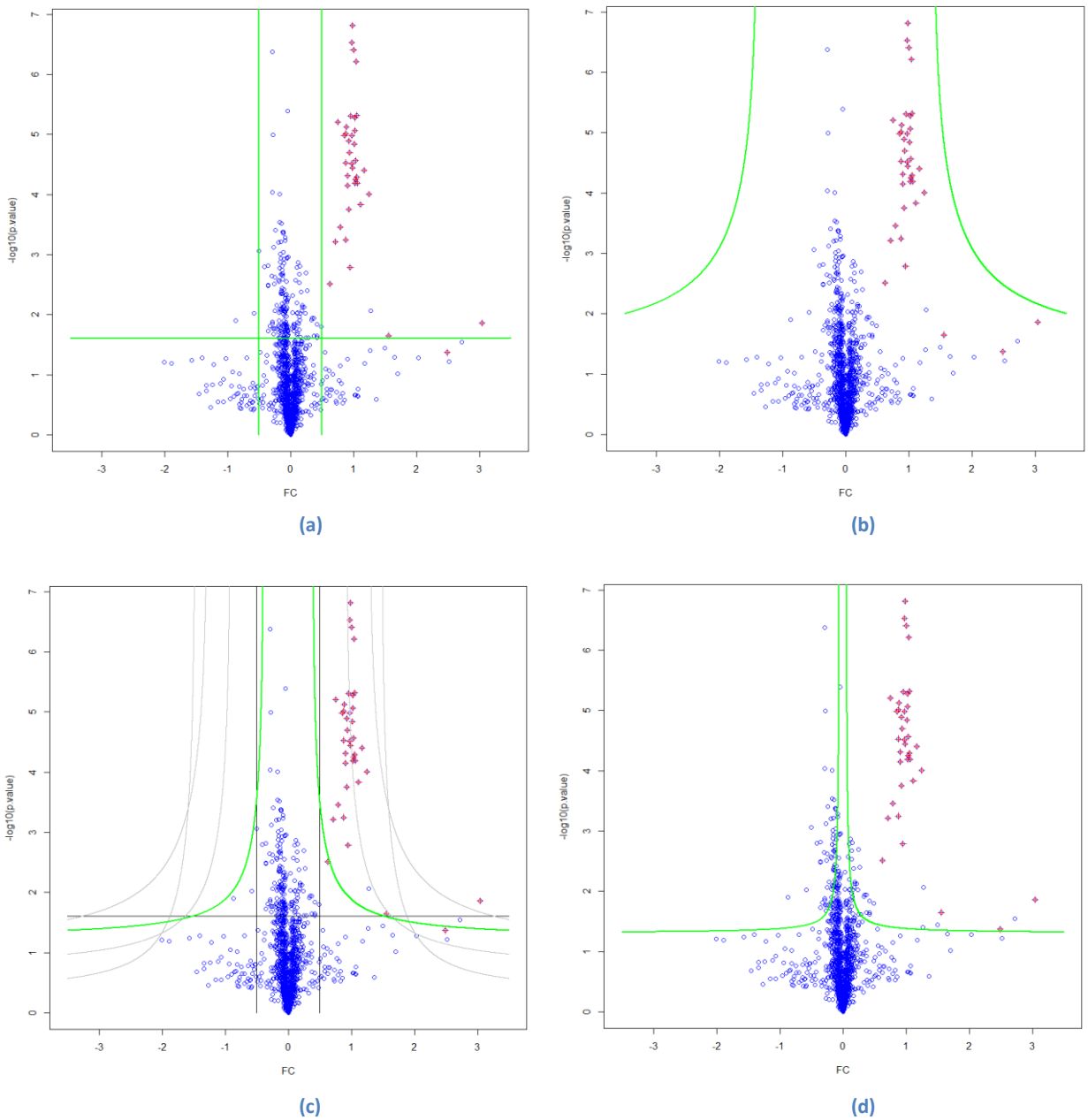

Figure 1: The volcano-plot of dataset LFQRatio2 (red crosses depict UPS1 proteins); (a) horizontal (p<2.5) and vertical (|*FC*|>0.5) thresholds; (b) smooth curve threshold, with s_0=0.5 and α=97.5%; (c) various parameter tuning are tried to mimic the selection of Fig. 1a: one ends up with s_0=0.15 and α=97%; (d) the automatic tuning according to [8].

Nonetheless, it is possible to perform iterative random guesses to converge toward some values of t_a and s_0 that correspond to our expectations, as illustrated on Fig. 1c. However, to achieve so, one has modified the SAM-test tuning and the corresponding *p*-values, which are not displayed on the volcano-plot (recall that the *p*-values of the Y-axis are those of a classical *t*-test). As already mentioned, this

distortion has no statistical justification and can lead to spurious biological conclusions. For comparisons, Fig. 1d displays the same curves, yet, with the automatic tuning advised by [8]; which is not expected to filter proteins on their *FC*.

Discussion: Although some proteomics studies were precursor in the use of statistical tools to analyze expression patterns [18][19], numerous proteomics studies still only rely on *FC* to claim that a protein is differentially abundant between compared conditions [20]. When statisticians are involved in proteomic studies, they naturally claim that *FC* alone is not a meaningful criterion to select differentially abundant proteins. However, selecting them only on *p*-values is not sufficient since it can lead to numerous false positives (proteins with very low *FC* that are unlikely to be biological relevant, or that appear as differentially abundant by chance, due to very low fluctuations between the replicates). This confrontation between the proteomists' needs and the statistical guidelines has led to the use of double thresholding. Basically, if statistical test and *FC* thresholding are not sufficient on their own, it is mainly because they rely on complementary viewpoints: On one hand, the vertical threshold on the *FC* corresponds to a practitioner's choice. Depending on the aim of the study, his/her knowledge of the protein abundance across various biological samples, etc., he/she may decide to be more or less stringent, and to focus on proteins with a great enough *FC*. These are proteomist's motivations that cannot be judged through the eye of statistics. On the other hand, the horizontal threshold on the *p*-value is tuned according to statistical guidelines: If the false discoveries are likely to be too numerous, the threshold should be increased, and if a greater number of false discoveries remains acceptable from a statistical viewpoint, one may decrease it. Finally, it is very important to understand that despite their similarities, these thresholds correspond to different expertise (respectively, proteomics and statistics), so that, they should not be mixed, swapped or confused, as they tend to be with a smooth curved threshold.

Under this light, the wrong tuning of s_0 to a too large value appears as a way to make *FC* thresholding looks like a statistical test (as it provides *p*-values), while practically circumventing the statistical guidelines (that requires to account for the variance of the protein abundances through the test). Concretely, its interest is to provide very tunable parametric curves to allow the user to precisely select the proteins he/she wants; and to derive *p*-values thanks to an artificial relationship between these curves and the SAM-test. As the method makes it possible to select exactly the "good proteins" (here the UPS1 ones), one supposes the procedure is statistically valid. However, only the correctness of the *p*-values is at stake, not the quality of the protein selection: Otherwise, it would be easier to propose the practitioner to draw a line surrounding the proteins of interest, and to randomly generate *p*-values. Although clearly dubious, it would not be arguably worse than applying unbounded distortions to real *p*-values, as with SAM-test misuse. In addition, such disguise of *FC* thresholding into statistical test gives the insidious impression that the filtering on the *FC* should have a statistical basis, for the practitioner's expertise is too subjective to rely on. On the contrary, this is typically one of the subjects that are beyond statisticians' expertise, and for which it may be dangerous to rely on theirs, rather than on the proteomics experts' one.

So, what to do instead? Finally, as practical guidelines, we advise to go back to the original use of volcano-plot, with independent vertical/horizontal thresholds corresponding to these complementary views. **The vertical thresholds should be performed first**, on the exclusive basis of the proteomist's expertise: the aim is only to filter out proteins with *FC* that is so small that they cannot be considered as biologically/analytically relevant, regardless a possible excellent *p*-value. In other words, one should not

define the *FC* cutoff to the highest possible value, so as to discard as many proteins as possible (this is the concern of the next coming "horizontal threshold"); but on the contrary, to the minimum value below which it is not even necessary for the biologist to have a deeper look. **Afterward, the statistical thresholding can be performed**. To do so, one defines a maximum authorized *p*-value, generally, on the basis of the expected false discovery rate associated to the set of proteins considered as differentially abundant.

At this point, it is important to note that the *p*-values for the volcano-plot can have previously been computed with any statistical test that is used according to its published uses, as depicted by its original authors. In this article, we certainly do not claim "SAM-test is bad"; it can be efficiently used, or not, depending on its tuning. Notably, in the case of automatic tuning (according to [8] recommendations) the practical advantages of the regularization largely counterbalance the minor distortions of the *p*-values. Moreover, beside SAM method, other options are possible, such as for instance, the empirical Bayesian approach of Limma [21][22]. Interestingly, thanks to the function `treat()` available in the corresponding R package, it is possible to incorporate a user-defined *FC* threshold in the null hypothesis definition , so as to account for it in a statistically valid way (see [23]).

Of course, it is possible to wonder on the best statistical test among all those available in the literature, and to compare them. However, it was not our goal here, as we only wanted to pinpoint that (1) any good test can lead to poor results if inadequately used; (2) statistical tests, whatever their quality, cannot provide a justification to some of the proteomists' choices, that mainly rely on their expertise, such as for instance, the *FC* cut-offs.

Acknowledgements

This work was funded through the French National Agency for Research grants ANR-10-INBS-08 (ProFI project, "Infrastructures Nationales en Biologie et Santé"; "Investissements d'Avenir"call) and ANR-13-BSV2-0012 (RNAGermSilence project).

References

[1] 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.

[2] S. Wieczorek, F. Combes, C. Lazar, Q. Giai-Gianetto, L. Gatto, A. Dorffer, A.-M. Hesse, Y. Coute, M. Ferro, C. Bruley, T. Burger. "DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics", under review, 2016. https://www.bioconductor.org/packages/3.3/bioc/html/Prostar.html

[3] 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.

[4] Andlauer TFM, Scholz-Kornehl S, Tian R, et al. Drep-2 is a novel synaptic protein important for learning and memory. Griffith LC, ed. eLife. 2014;3:e03895. doi:10.7554/eLife.03895.

[5] Hubner NC, Bird AW, Cox J, et al. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. The Journal of Cell Biology. 2010; 189(4):739-754. doi:10.1083/jcb.200911091.

[6] Hornburg D, Drepper C, Butter F, Meissner F, Sendtner M, Mann M. Deep Proteomic Evaluation of Primary and Cell Line Motoneuron Disease Models Delineates Major Differences in Neuronal Characteristics. Molecular & Cellular Proteomics : MCP. 2014;13(12):3410-3420. doi:10.1074/mcp.M113.037291.

[7] Cezary Smaczniak, Na Li, Sjef Boeren, Twan America, Walter van Dongen, Soenita S Goerdayal, Sacco de Vries, Gerco C Angenent, Kerstin Kaufmann. Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues, Nature protocols, 2012.

[8] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(9):5116-5121. doi:10.1073/pnas.091062498.

[9] Tikhonov AN. On the stability of inverse problems, Dokl. Akad. Nauk SSSR, 39, No. 5, 195-198, 1943.

[10] Wu B. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics*, *21*(8), 1565-1571. 2005

[11] Efron B, Tibshirani R, Storey JD, & Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, *96*(456), 1151-1160. 2001.

[12] Schwender H. *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*. R package version 1.44.0. 2012

[13] Giai Gianetto Q, Combes F, Ramus C, Bruley C, Couté Y and Burger T. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. Proteomics, 16: 29–32. 2016; doi:10.1002/pmic.201500189

[14] Lundby A, Andersen MN, Steffensen AB, Horn H, Kelstrup CD, Francavilla C, ... and Olsen JV. In vivo phosphoproteomics analysis reveals the cardiac targets of β-adrenergic receptor signaling. *Sci. Signal.*,*6*(278), rs11-rs11. 2013.

[15] Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. Molecular & Cellular Proteomics : MCP. 2012;11(3):M111.014050. doi:10.1074/mcp.M111.014050.

[16] Smits AH, Jansen PWTC, Poser I, Hyman AA, Vermeulen M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. Nucleic Acids Research. 2013;41(1):e28. doi:10.1093/nar/gks941.

[17] Lin, D., Shkedy, Z., Burzykowski, T., Ion, R., Göhlmann, H. W. H., Bondt, A. D., ... & Bijnens, L. (2008). An Investigation on Performance of Significance Analysis of Microarray (SAM) for the Comparisons of Several Treatments with one Control in the Presence of Small-variance Genes. Biometrical Journal,50(5), 801-823.

[18] Tarroux, P. (1983). Analysis of protein patterns during differentiation using 2-D electrophoresis and computer multidimensional classification. *Electrophoresis*,*4*(1), 63-70.

[19] Anderson, N. L., Hofmann, J. P., Gemmell, A., & Taylor, J. (1984). Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clinical chemistry*, *30*(12), 2031-2036.

[20] Ting, L., Cowley, M. J., Hoon, S. L., Guilhaus, M., Raftery, M. J., & Cavicchioli, R. (2009). Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular & Cellular Proteomics*, *8*(10), 2227-2242.

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

[22] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47. doi:10.1093/nar/gkv007.

[23] McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25(6):765-771. doi:10.1093/bioinformatics/btp053.