Technical Brief

Calibration Plot for Proteomics (CP4P): A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments

Quentin Giai Gianetto[1,3,4], Florence Combes[1,3,4], Claire Ramus[1,2,3,4], Christophe Bruley[1,3,4], Yohann Couté[1,3,4], Thomas Burger[1,2,3,4]

[1] Univ. Grenoble Alpes, iRTSV-BGE, F-38000 Grenoble, France. [2] CNRS, iRTSV-BGE, F-38000 Grenoble, France. [3] CEA, iRTSV-BGE, F-38000 Grenoble, France. [4] INSERM, BGE, F-38000 Grenoble, France.

November 4, 2015

Abstract: In mass-spectrometry based quantitative proteomics, the false discovery rate control (*i.e.* the limitation of the number of proteins which are wrongly claimed as differentially abundant between several conditions) is a major post-analysis step. It is classically achieved thanks to a specific statistical procedure which computes the *adjusted p-values* of the putative differentially abundant proteins. Unfortunately, such adjustment is conservative only if the p-values are *well-calibrated*; the false discovery control being spuriously underestimated otherwise. However, well-calibration is a property that can be violated in some practical cases. To overcome this limitation, we propose a graphical method to straightforwardly and visually assess the p-value well-calibration, as well as the `R` codes to embed it in any pipeline.
Keywords: False discovery rate; Relative quantification experiments; Statistical significance.

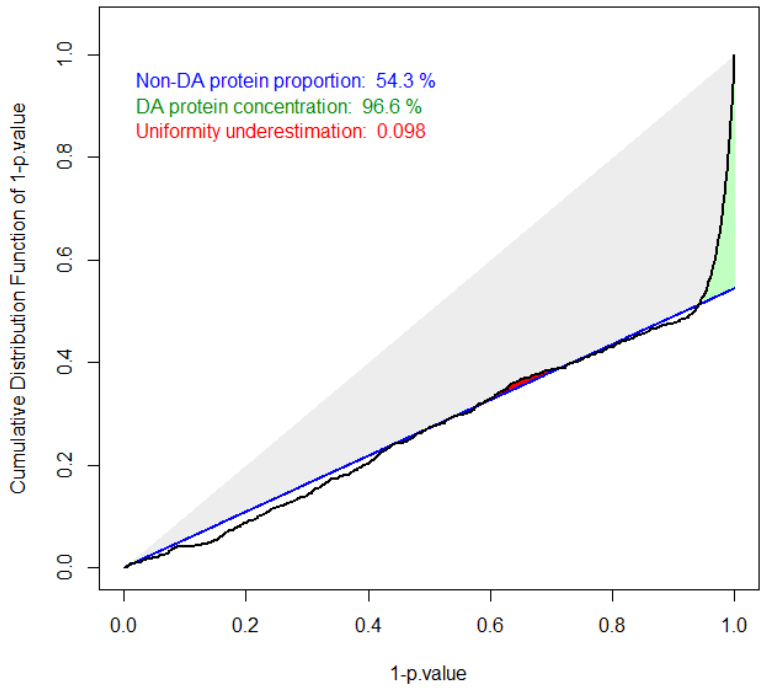

Figure 1: Typical graphical output of the `calibration.plot()` function on a dataset with well-calibrated p-values.

In high-throughput proteomics, relative bottom-up quantification refers to the search of differentially abundant proteins between at least two conditions. First, several replicated samples for each condition are digested with trypsin, then analyzed by liquid chromatography and tandem mass spectrometry. Second, the output of these analyses is processed by bioinformatics tools, so as to identify peptides, to aggregate them into proteins, and to provide an abundance value for each replicate and each protein [1]. Third, a statistical test is performed to find proteins which are *differentially abundant* (DA) in one of these conditions (see supplemental material). The test outcomes do not directly separate DA proteins from non-DA proteins but provide instead a list of p-values that are related to a protein each, and that must be interpreted and filtered: there are sometimes proteins that are not DA while having low p-values, possibly leading to false discoveries. This is why FDR (false discovery rate) control is required afterward.

Introduced by Benjamini and Hochberg [2], FDR originally referred to the estimation of the expectation of the proportion of false discoveries in a list of putative discoveries. Since then, numerous other improvements to the method were proposed [3, 4, 5], as well as variations on the statistical quantity to control [6, 7]. However, to date, it is possible to summarize a general pattern common to the most used FDR control procedures: (i) compute p_i (the p-value of the i^{th} protein in a protein list of length m) ; (ii) reorder the protein list so that $p_{(1)}$ is the smallest p-value and $p_{(m)}$ the greatest; (iii) transform each $p_{(i)}$ into $p^*_{(i)}$ the so-called *adjusted p-value* (or *q-value* [6]) which corresponds to the smallest FDR at which the corresponding protein will be concluded DA; (iv) cut the list to $n \leq m$ so that $p^*_{(n)}$ corresponds to the desired FDR level.

The theoretical foundations of FDR require that the raw p-values respect some specific assumptions [8]: there is an unknown proportion π_0 of non-DA proteins, the p-values of which are uniformly distributed on the [0,1] interval, while the remaining p-values (corresponding to DA proteins) are concentrated nearby zero. If this strong mathematical hypothesis is violated (in such a case, the p-values are said *badly-calibrated*), the FDR control may be spurious, possibly leading to false biological conclusions. If the p-values are badly calibrated, in theory, it is possible to rely on few specific FDR control procedures that require less restrictive mathematical assumptions: The first one is the Benjamini-Yekutieli procedure [4]. However, in most of the cases, it is so conservative that it drastically reduces the number of proteins that can be assumed DA; so that practitioners are reluctant to use it, whatever its statistical robustness. The second one is to rely on some permutation-based procedure [7, 9]. However, for the number of permutations to be high enough to ensure reliable results, it is mandatory to have more samples than in usual proteomics experiment, making them hardly compliant with everyday proteomics constraints. As a result, it is most important to check, at least visually, that the p-values are well-calibrated.

This calibration issue has long been known in statistics [10], in genome-wide association studies [11, 12], or in meta-analyses [13]. Yet, to the best of our knowledge, it has not penetrated the proteomics community so far. This is a concern as it is frequent to remove entries of the protein list for sensible reasons (*e.g.*, proteins with low fold-change, or proteins identified with weak evidence, etc.). Such a filtering is guided by the practitioner's motivations, and it has no reason to operate uniformly over the range of p-values. Thus, it may involve a change in their distribution, leading to bad-calibration cases. This is why, we proposed a set of `R` functions, embedded in a dedicated package named `CP4P (Calibration Plot for Proteomics)` [14], which allows to visually assess the well-calibration of the p-values.

The function `calibration.plot()` of the CP4P package takes as input a vector of previously computed p-values. As output, it provides a graph similar to Fig. 1, which displays (black curve) the cumulative distribution function of $1 - p_i$ ($i \in [1, m]$) as a function of $1 - p_i$ such as advocated in [10]. As it clearly appears, the curve starts from point [0,0], and is then roughly linear indicating that the non-DA proteins have p-values that are roughly uniformly distributed. On the other hand, the curve becomes very peaky nearby the $[0.9, 1]$ interval, indicating that

there is an important concentration of small p-values, most likely corresponding to DA proteins.

Moreover, a blue line is displayed. It is expected to have the same trend as the linear part of the black curve, as illustrated on Fig. 1. The slope of this blue line corresponds to an estimation of the proportion of non-DA proteins (classically noted π_0), which is indicated as *non-DA protein proportion* (in blue too). Concretely, its equation reads $y = \pi_0 x$.

The area A between the right hand side peak of the black curve and the blue line is colored in green. This area is important: it depicts the extent to which the set of DA proteins have different p-values than other proteins, and consequently, the extent to which they can be discriminated on the basis of a good FDR threshold. The thinner this area, the better, as it amounts to having DA proteins with p-values distinctly smaller than the others (so that the false non-discovery rate [15] is expected to be smaller). To propose a quantitative estimation of the quality of the p-value distribution in relationship with this green area, we derived the *DA protein concentration* measure that reads $1 - A/T$, where T is the gray triangle area (by construction, $T = (1 - \pi_0)/2$). This concentration appears in green in the top left corner of the plot; intuitively, the closer to 100%, the better: For instance, on Fig. 1, the concentration is nearly perfect (96.8%), while on Fig. 3(right), it is too low (52.7%) to expect a clear discrimination of the DA proteins.

Finally, the *uniformity underestimation* (in red) corresponds to the area where the black curve is above the blue line apart from the peak at the left hand side (DA protein peak). To get a conservative adjustment of the p-values (so that it does not under-estimate the FDR), the black curve has to remain below the blue line (see [8], or [2] for justifications). In the ideal case, the left hand side of the black curve always remains below the blue one and the uniformity underestimation is null. However, in practice, as long as the uniformity underestimation remains small (below a guesstimate of 0.5), p-values can be adjusted.

As explained, the blue line is of prime importance for the visual assessment of the p-value distribution. However, its slope, reflecting π_0 needs to be estimated since it is unknown. To do so, the Pounds estimator proposed in [16] is used by default: in addition to rely on solid theoretical foundations, it appears to provide good estimations on most of our experiments. However, as any estimator, it is possible to exhibit situations where it is inaccurate. For this reason, one may want to use other state-of-the-art estimation methods instead. Concretely, this is implemented in `calibration.plot()` with a second optional argument, which can take several values:

- A value x between 0 and 1, corresponding to the freely tuned proportion of non-DA proteins, for cases where the pratitionner knows the precise content of the sample.
- The name of an estimation method among: "pounds" [16] (default), "st.boot" [11], "st.spline" [17], "langaas" [18], "jiang" [19], "histo" [20], "abh" [3], "slim" [21].
- "ALL": A different line for the eight aforementioned methods is displayed so that the practitioner chooses on his/her own, the most adapted one. This is illustrated on Fig. 2 from a controlled dataset[1] containing only 38 DA proteins among 1481 proteins ($\pi_0 = 97.4\%$), so that the Pounds estimator is rather inaccurate ($\approx 85\%$). This points out the interest of choosing alternative methods (such as the "abh" in this case).

Conveniently, `calibration.plot()` comes along with `adjust.p()`, a function which binds the classical p-value adjustment method [22] with an extra argument that allows choosing the estimation method for π_0. It is possible to give its numerical value (by default, the parameter is tuned to $\pi_0 = 1$ to enforce maximal conservativeness, see [2]), or the name of any of the eight aforementioned methods. Finally, if the practitioner is not satisfied by any π_0 estimation method, while expecting a precise FDR level, it remains possible to use the "bky" method [5]

[1] The Sigma UPS1 equimolar mixture, spiked in a yeast background. This dataset can be found on the ProteomeXchange repository with the identifier PXD002370.

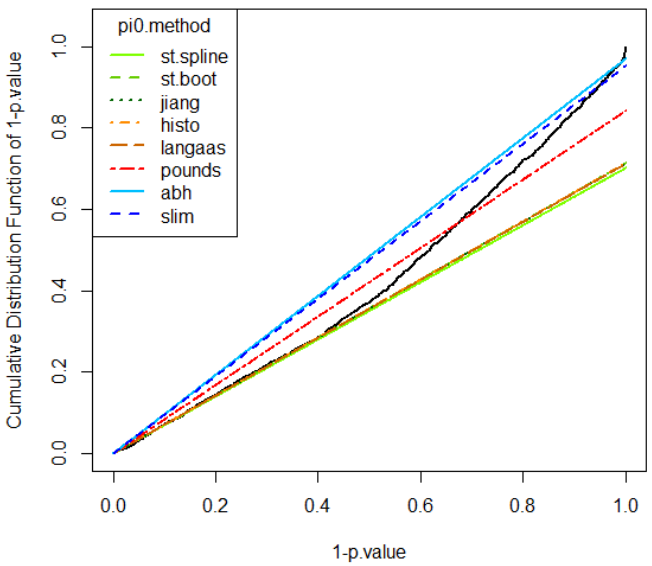

Figure 2: Illustration of the "ALL" optional argument on a real dataset where the default π_0 estimator is inaccurate.

which dynamically tunes π_0. Despite the drawback of linking π_0 to an arbitrary FDR level, this last option remains interesting as this procedure is theoretically proven to be conservative. Note that in this package, we have promoted FDR control procedures of the BH family. However, other methods exists, that control a slightly different quantity, as described in [6]. Once the well-calibration is assessed, any practitioner is free to use another method, regardless our implementations.

For better understanding, Fig. 3 displays two counter-examples: In the first example (Fig. 3-left), the p-values of a simulated dataset[2] are not correctly distributed (i.e. badly calibrated): π_0 makes sense and the *DA protein concentration* is close to 1. However, the *uniformity underestimation* is clearly too high: there are too many high p-values (close to 1) so that the assumption of uniformity is not respected and the FDR control underestimates the true proportion of false discoveries. On Fig. 3-right, the same dataset as in Fig. 2 has been considered and the "st.boot" method has been applied. As a result, π_0 is badly estimated: even if almost no red area shows up, there is a too big green area according to the expectations. Indeed, π_0 should be set to a value really close to 100% (the real value being 97.4% in this example). This example is detailled in the CP4P tutorial (supplemental data), along with a calibration study of the proteomics experiments described in [23].

To conclude, FDR control procedures rely on strong assumptions that must be checked before proper application. Thanks to CP4P, proteomicians can visually assess the well-calibration of the p-values and estimate the proportion of non-DA proteins in relative quantification experiments. CP4P allows processing any quantitative datasets, regardless the experimental design (nature and hierarchy of replicates, number of conditions, etc.), it is easy-to-use and it can be embedded in any bioinformatics pipeline via the CRAN [14].

Acknowledgments

This work was supported by ANR-10-INBS-08 (ProFI project, "Infrastructures Nationales en Biologie et Santé", "Investissements d'Avenir"), ANR-13-BSV2-0012 (RNAGermSilence project), and the Prospectom project (Mastodons 2012 CNRS challenge). The proteomics data have been deposited to the ProteomeXchange Consortium [24] via the PRIDE partner repository with the dataset identifier PXD002370.

[2] 1000 p-values are generated using a mixture of two equally weighted Beta distributions, with parameters $\alpha_1 = 1$, $\beta_1 = 20$, $\alpha_2 = 5$ and $\beta_2 = 1$.

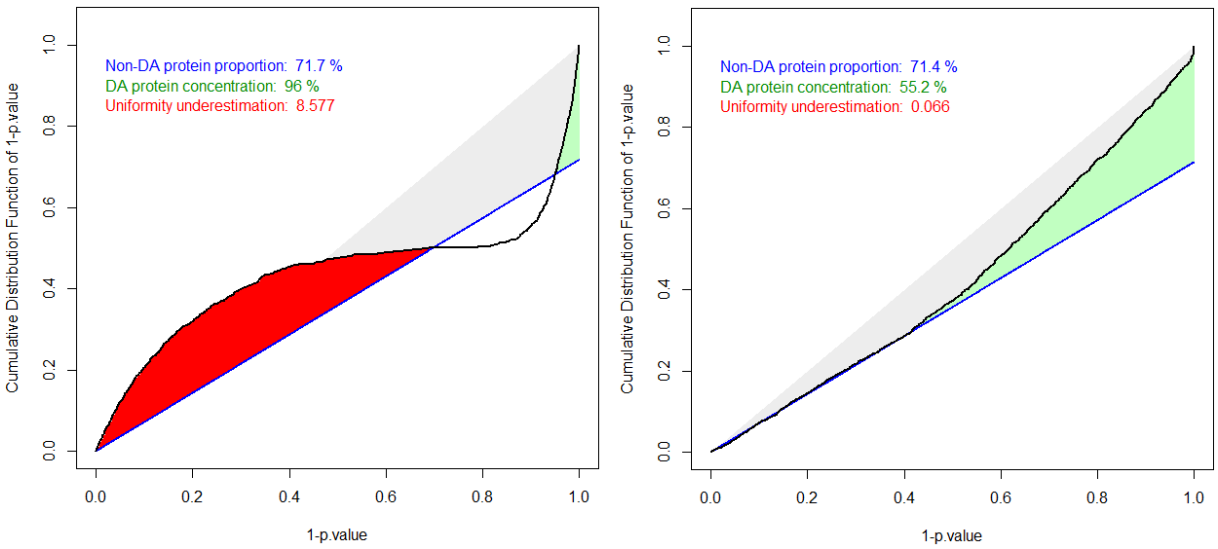

Figure 3: (Left) A simulated example of badly-calibrated p-values; (Right) A real dataset where π_0 is inaccurately estimated.

References

[1] Sven Nahnsen, Chris Bielow, Knut Reinert, and Oliver Kohlbacher. Tools for label-free peptide quantification. *Molecular & Cellular Proteomics*, 12(3):549–556, 2013.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[3] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[5] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[6] John D Storey. False discovery rates. In *International encyclopedia of statistical science*, pages 504–508. Springer Berlin Heidelberg, 2011.

[7] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[8] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

[9] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1):171–196, 1999.

[10] Tore Schweder and Emil Spjøtvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.

[11] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[12] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[13] http://cran.r-project.org/web/packages/metap.

[14] http://cran.r-project.org/web/packages/cp4p.

[15] Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[16] Stan Pounds and Cheng Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987, 2006.

[17] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[18] Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):555–572, 2005.

[19] Hongmei Jiang and RW Doerge. Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer informatics*, 6:25, 2008.

[20] Dan Nettleton, JT Gene Hwang, Rico A Caldo, and Roger P Wise. Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337–356, 2006.

[21] Hong-Qiang Wang, Lindsey K Tuominen, and Chung-Jui Tsai. Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, 27(2):225–231, 2011.

[22] Yoav Benjamini, Effi Kenigsberg, Anat Reiner, and Daniel Yekutieli. Fdr adjustments of microarray experiments (fdr-ame). *R package version 1.38.0.*, 2013.

[23] Yacine Bounab, Anne-Marie Hesse, Bruno Iannascoli, Luca Grieco, Yohann Coute, Anna Niarakis, Romain Roncagalli, Eunkyung Lie, Kong-Peng Lam, Caroline Demangel, Denis Thieffry, Jerome Garin, Bernard Malissen, and Marc Daeron. Proteomic Analysis of the SH2 Domain-containing Leukocyte Protein of 76 kDa (SLP76) Interactome in Resting and Activated Primary Mast Cells. *Molecular & Cellular Proteomics*, 13(2):678, 2014.

[24] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. Proteomexchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3):223–226, 2014.