BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 21 2013, pages 2729–2734
doi:10.1093/bioinformatics/btt464

Gene expression

Advance Access publication August 27, 2013

Identifying differentially expressed proteins in two-dimensional electrophoresis experiments: inputs from transcriptomics statistical tools

Sébastien Artigaud*, Olivier Gauthier and Vianney Pichereau

Laboratoire des Sciences de l'Environnement Marin, LEMAR UMR 6539 CNRS/UBO/IRD/Ifremer, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale, 29280 Plouzané, France

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Background: Two-dimensional electrophoresis is a crucial method in proteomics that allows the characterization of proteins' function and expression. This usually implies the identification of proteins that are differentially expressed between two contrasting conditions, for example, healthy versus diseased in human proteomics biomarker discovery and stressful conditions versus control in animal experimentation. The statistical procedures that lead to such identifications are critical steps in the 2-DE analysis workflow. They include a normalization step and a test and probability correction for multiple testing. Statistical issues caused by the high dimensionality of the data and large-scale multiple testing have been a more active topic in transcriptomics than proteomics, especially in microarray analysis. We thus propose to adapt innovative statistical tools developed for microarray analysis and incorporate them in the 2-DE analysis pipeline.

Results: In this article, we evaluate the performance of different normalization procedures, different statistical tests and false discovery rate calculation methods with both real and simulated datasets. We demonstrate that the use of statistical procedures adapted from microarrays lead to notable increase in power as well as a minimization of false-positive discovery rate. More specifically, we obtained the best results in terms of reliability and sensibility when using the 'moderate t-test' from Smyth in association with classic false discovery rate from Benjamini and Hochberg.

Availability: The methods discussed are freely available in the 'prot2D' open source R-package from Bioconductor (http://www.bioconductor.org//) under the terms of the GNU General Public License (version 2 or later).

Contact: sebastien.artigaud@univ-brest.fr or sebastien.artigaud@gmx.com

Received on May 8, 2013; revised on July 3, 2013; accepted on August 5, 2013

1 INTRODUCTION

Comparative proteomics based on 2D-polyacrylamide gel electrophoresis aims at identifying significant biologically relevant proteins. In most cases, protein samples from two conditions are compared, e.g. healthy versus diseased in human proteomics biomarker discovery (Kim *et al*., 2004), or stressful conditions versus control in animal experimentation (Tomanek *et al*., 2011). Besides the laboratory bench, informatics takes a large place in such experiments with digitalization of gels, spots detection and matching, evaluation of spots' volumes and identification of differentially expressed proteins.

A large variety of software designed for the analysis of 2D digitized images exist, both commercial (e.g. Delta2D from Decodon, SameSpot Progenesis from Nonlinear dynamics, PDQuest from Bio-rad laboratories) and custom researcher-developed (e.g. Pinnacle from Morris *et al*., 2010; RegStatGel from Li and Seillier-Moiseiwitsch, 2011). These programs perform three main tasks: identification of spots, matching of spots from different gels and evaluation of spots' volumes. Each program has its own way to perform these steps; therefore, software choice has a great impact on the result of the analysis (Millioni *et al*., 2012; Morris *et al*., 2010; Stessl *et al*., 2009; Wheelock and Buckpitt, 2005). A normalization step is also needed to remove the systemic variation before data analysis. An issue with commercial software is the lack of information concerning this critical step as well as the inability to customize the normalization procedure. Ultimately, once spots' volumes have been estimated and normalized across gels, proteins that are differentially expressed between groups are identified with software-specific methods.

Normalization and identification are also essential steps in 2-DE-based proteomics analysis and, as compared with other critical steps of the 2-DE procedure (e.g. protein extraction, staining, image analysis), are less studied and under-reviewed in the literature. On the other hand, the same issues have been widely studied and reviewed for transcriptomic analysis, especially for microarray studies. As pointed out by some authors, the underlying statistical issues are similar between microarray- and 2-DE analyses. A number of the statistical tools developed for microarray analysis have begun to be incorporated in the 2-DE toolbox. More specifically, advances in normalization have been adapted to Difference Gel Electrophoresis (DIGE) analysis (Fodor *et al*., 2005; Miecznikowski *et al*., 2011). Statistical tests developed for microarrays have also been evaluated (Fodor *et al*., 2005; Meunier *et al*., 2005). Finally, numerous authors highlight the need of correction for multiple testing, such as false discovery rate (FDR), but few actually evaluate the performance of the different procedures (Chang *et al*., 2004; Karp *et al*., 2007; Morris, 2012).

The aim of the present work is to evaluate different methods, initially developed for microarray analysis, to identify differential

*To whom correspondence should be addressed.

spots in 2-DE experiments. To compare these methods, real datasets as well as simulated data were used. We also compare methods for normalization of volumes' data before statistical analysis. By evaluating the key steps of the analysis workflow, from normalization to identification of differentially expressed spots, our motivation is to provide a simple and adapted workflow to non-statistics expert biochemists. All the methods discussed are available as a free, open-source R package (R Core Team, 2012) with highly customizable options.

2 METHODS

2.1 Datasets description

2.1.1 Platichthys flesus *liver dataset* The *Platichthys flesus* liver dataset (Pfl DS)corresponded to a 2-DE-based comparison of liver proteomes of two populations of *Platichthys flesus* living in contrasted estuaries, i.e. the Seine estuary, known to be highly polluted, and the Canche, which is often considered as a pristine area. The whole study was previously published in Galland *et al.* (2013). It included analyzing the liver proteomes of four juvenile individuals from each estuary.

2.1.2 Pecten maximus *gills dataset* The *Pecten maximus* gills dataset (Pmg DS) is issued from a 2-DE experiment performed on proteins from *Pecten maximus* gills subjected to a temperature challenge. Briefly, scallops were kept in two separate tanks at 16.4°C (±0.3°C) (Huber *et al.*, 2002) and fed *ad libitum* for 2 weeks. One tank was then heated at 1°C per day and allowed to reach 27°C ('hot condition'), whereas the other tank was kept at 16.5°C ('control'). At the end of the experiment, six animals per condition were sampled. Gills were snap-frozen in liquid nitrogen and kept at −80°C until protein extraction. Gills were crushed (using Retsch® MM400) device kept frozen using liquid nitrogen. Hundred milligrams of the obtained powder was homogenized in 100 mM Tris–HCl (pH 6.8), centrifuged (4°C, 50 000 *g*, 5 min) and supernatants were pipetted in other tubes. Protease inhibitor mix (GE Healthcare) was then added and nucleic acids were removed (nuclease mix, GE Healthcare, following manufacturer's instructions). Samples were precipitated overnight at −20°C using TCA 20% (1/1:v/v, overnight). After centrifugation (4°C, 20 000 *g*, 30 min), pellets were washed with acetone 70% and 0.1% DTT and re-suspended in urea/thiourea buffer (2 M thiourea, 7 M urea, 4% CHAPS, 1% DTT) containing 1% IPG (pH 3–10, GE Healthcare). Protein concentration was determined using a modified Bradford assay12, and all samples were adjusted to 400 mg of proteins in 250 µl. Electrophoresis and staining procedures were performed as described in Galland *et al.* (2013).

2.2 Preprocessing of data

2.2.1 Image Analysis All gels were digitized using a transparency scanner (Epson Perfection V700) in gray scale with 16-bit depth and a resolution of 600 dpi. Images were aligned and spots were detected and quantified with the Progenesis SameSpots software (Nonlinear dynamics, v.3.3) using the automated algorithm. All detected spots were manually carefully checked and artifact spots were removed. Datasets were exported as raw values in the form of a matrix of volume data X_{ij} with spots i as rows and gels j as columns. In all, 611 spots were identified in the Pfl dataset from Galland *et al.* (2013) with four replicates in two conditions (611 × 8 matrix), and 766 spots were identified in the Pmg dataset with six replicates per condition (766 × 12 matrix).

2.3 Visualization and normalization of datasets

Dudoit *et al.* (2002) proposed a method for visualization of artifacts in microarray datasets, called the MA-plot, which was transposed for proteomics data as the ratio-intensity plot (Meunier *et al.*, 2005; R-I plot). It consists in plotting the intensity $\log_2$-ratio (R) against mean $\log_{10}$ intensity (I):

$$R = \log_2 \frac{mean(V_{Cond2})}{mean(V_{Cond1})}$$

$$I = \log_{10}(mean(V_{Cond2}) \times (mean(V_{Cond1}))$$

where V_{Cond1} and V_{Cond2} are spots' volumes for conditions 1 and 2, respectively.

R-I plots allow to directly visualize artifacts in the original dataset as well as the effects of normalization (Fig. 1). Artifacts were already described for Coomassie blue-stained 2-DE gels, but have also been reported in DIGE, SYPRO ruby and silver-stained experiments. In the original datasets (Fig. 1A and B), the cloud of points seems to be off-centered, especially for low-intensity spots, either down-shifted in Pmg DS or up-shifted in Pfl DS.

Among the variety of methods for normalization recently proposed in the literature, both for proteomics and microarray experiments (Podrabsky and Somero, 2004; Quackenbush, 2002; Yang *et al.*, 2002), two widely used methods are compared, the 'Variance Stabilizing Normalization' (VSN; Huber *et al.*, 2002) and the 'Quantile Normalization' (Qt; Bolstad *et al.*, 2003). The principle of the 'quantile normalization' is to set each quantile of each column (i.e. the spots' volume data of each gels) to the mean of that quantile across gels. The intention is to make all the normalized columns have the same empirical distribution. Whereas the VSN method relies on a transformation h of the intensities, of the parametric form $h(x) = arsinh(a + bx)$ .The parameters of h are estimated with a robust variant of maximum-likelihood estimation. The R-I plots of normalized data (Fig. 1) show that both methods center the data around a log ratio of 0. Nevertheless, for low values of intensities the VSN normalized data (Fig. 1E and F) seem to be less efficient to recenter the cloud of points.

2.4 False discovery rate and significance

Transcriptomics and 2-DE experiments analyses have in common that they require a variant of a t statistic that is suitable for high-dimensional data and large-scale multiple testing. For this purpose, in the past few years, various test procedures have been suggested. We decided to compare a simple method (the classical Student's t-test), two tests especially

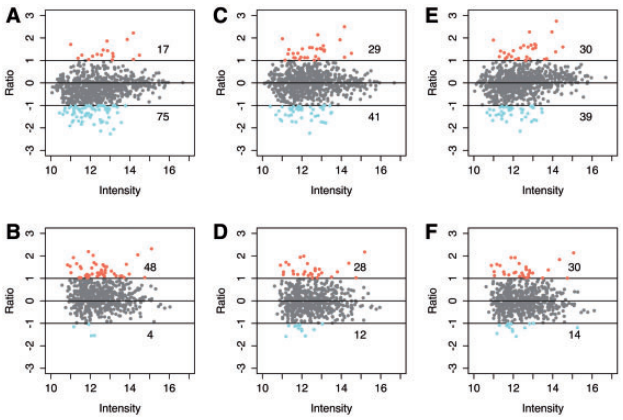

Fig. 1. Comparison of the effect of normalization procedures using R-I plots. Values on the plots are the number of spots with a ratio >1 (upper values) or <−1 (lower values). (**A**) Pmg DS without normalization. (**B**) Pfl DS without normalization. (**C**) Pmg DS with Qt. (**D**) Pfl DS with Qt. (**E**) Pmg DS with VSN. (**F**) Pfl DS with VSN

modified for microarray analysis (Efron's *t*-test; Efron *et al*., 2001), the modified *t*-test used in significance analysis for microarray (SAM; Tusher *et al*., 2001) and two methods that take advantage of hierarchical Bayes methods for estimation of the variance across genes: the 'moderate *t*-test' from Smyth (2004) and the 'Shrinkage *t*' statistic test from Opgen-Rhein and Strimmer (2007).

Statistical tests allowing the identification of differentially expressed proteins must take into account a correction for multiple tests to avoid false conclusions. Testing for expression changes across hundreds or thousands of proteins with univariate test, such as Student's *t*-test, causes false positives to accumulate. To address this issue in multiple comparison procedure, Benjamini and Hochberg (1995) proposed a method for estimating and controlling the FDR, allowing for 'FDR-driven' decision of significance.

In proteomics, the FDR is usually defined as the ratio of the proteins not differentially expressed in reality but declared differentially expressed by the test, over the total number of proteins declared differentially expressed by the test [Table 1; $FDR = FP/(TP + FP)$].

From a statistical point of view, there are two types of FDR, the 'classic' tail area-based FDR (*Fdr* in Efron, 2008) and the local FDR (*fdr* in Efron, 2008). As it is easier to interpret for non-specialists, we focused our study on tail area-based FDR methods.

We decided to compare four different FDR estimators: (i) the classical FDR estimator of Benjamini and Hochberg, (ii) Strimmer's FDR (based on local FDR calculation; Strimmer, 2008a), (iii) the 'robust Fdr' estimator of Pounds and Cheng (2006) and (iv) the widely used FDR method known as 'q-value' defined by Storey (2002) and improved in 2004 (Storey *et al*., 2004).

To evaluate the impact of the normalization on the number of spots declared significant as well as the impact of the statistical test and mode of calculation for FDR, the following workflow was used for each dataset:

(1) Normalization of raw volume data with either VSN or quantile methods.
(2) Calculation of statistic value for normalized volume data using Student's *t*-test, Moderate *t*-test, SAM statistic, Efron's *t*-test and shrinkage *t*-test.
(3) Computation of *P*-values for each test based on a null distribution estimated using the fdrtool package (Strimmer, 2008b).
(4) FDR values were calculated using *P*-values as input with Benjamini and Hochberg ('BH'), Strimmer ('Stri'), Pounds and Cheng ('PC') or Storey ('Sto') procedures.
(5) Spots under a cut-off of 0.1 were declared significantly differentially expressed between the conditions.

2.5 Simulations study

To compare FDR and the responses of the different tests as well as the influence of the number of replicates, simulated data were used. Data were simulated based on parameter estimates of Pmg dataset, following these steps:

(1) Log_2 mean volumes from Pmg DS were computed for each spot.
(2) Means were used as input parameters to simulate a normal distribution (with no differential expression between conditions) for each spot with standard deviations computed as described by Smyth (2004).
(3) Ten percent of spots were randomly picked for introducing differential expression in both conditions (5% in each condition).

Briefly, in this hierarchical Bayesian model, the distribution is controlled by hyperparameters s_0^2 (estimator of the standard deviation) and d_0 (the degrees of freedom of the $\chi 2$ distribution used in the calculation of the distribution; see Smyth, 2004 for details). To simulate realistic data, we used $s_0^2 = 0.2$ and $d_0 = 3$.

Table 1. Output of statistical test versus reality of protein expression

	Declared non-significant	Declared significant
Protein *is not* differentially expressed	True negative	False positive
Protein *is* differentially expressed	False negative	True positive

3 RESULTS AND DISCUSSION

3.1 Results from Pmg and Pfl datasets

As seen in Table 2, both the moderate *t*-test and the classical Student's *t*-test seem to (i) be the most coherent among the compared tests and (ii) detect a greater number of differentially expressed spots. The effectiveness of the moderate *t*-test stems from its very definition: a modified *t*-test for which the standard errors have been moderated across spots, i.e. shrunk toward a common value, using a simple Bayesian model. As already demonstrated with microarray data, the smoothing of the standard errors increases the reliability of the test (Smyth, 2004).

Table 2. Spots declared significant (FDR < 0.1) for Pmg and Pfl DS with two methods for normalization: VSN and Qt. Four methods of FDR calculation are also compared: BH (Benjamini and Hochberg, 1995); Stri (Strimmer, 2008a); PC (Pounds and Cheng, 2006); Sto (Storey *et al*, 2004)

Normalization	FDR	Student's *t*-test	Moderate *t*-test	SAM's *t*-test	Efron's *t*-test	Shrinkage *t*-test
Pmg DS						
VSN	BH	4 (1)	7 (2)	4 (4)	3 (3)	1 (1)
	Stri	4 (1)	7 (2)	4 (4)	3 (3)	1 (1)
	PC	0	7 (2)	160 (15)	172 (15)	4 (4)
	Sto	4 (1)	7 (2)	4 (4)	3 (3)	1 (1)
Qt	BH	2 (1)	16 (3)	4 (4)	5 (5)	1 (1)
	Stri	3 (1)	20 (9)	4 (4)	5 (5)	1 (1)
	PC	2 (1)	16 (3)	0	85 (17)	0
	Sto	2 (1)	21 (9)	4 (4)	5 (5)	2 (2)
Pfl DS						
VSN	BH	9 (1)	14 (7)	1 (1)	1 (1)	1 (1)
	Stri	9 (1)	14 (7)	1 (1)	1 (1)	1 (1)
	PC	40 (9)	19 (8)	19 (16)	1 (1)	1 (1)
	Sto	10 (1)	14 (7)	1 (1)	1 (1)	1 (1)
Qt	BH	6 (0)	9 (5)	0	0	0
	Stri	6 (0)	9 (5)	0	0	0
	PC	6 (0)	9 (5)	0	0	0
	Sto	6 (0)	9 (5)	0	0	0

Note: Values in parentheses are the number of spots with an absolute log_2 ratio > 1.

Another striking fact is the inconstancy of Pounds and Cheng's 'robust Fdr' method throughout the results (e.g. from 0 spots detected for classic *t*-test to 172 for Efron's *t*-test for Pmg VSN data). This might reflect that 'robust Fdr' is not well adapted to 2D gel data, where the null proportion (the proportion of spots not differentially expressed) is high. Actually, the authors warned about the instability of the FDR estimates when this null proportion nears 1 (Pounds and Cheng, 2006). Putting aside 'robust Fdr' results, SAM's, Efron's and the shrinkage *t*-test are relatively consistent with both normalization methods but detect almost no significant proteins in the Pfl dataset. Finally, the standard Student *t*-test is consistent among all conditions but detected less significant proteins than the moderate *t*-test.

Concerning the normalization method, the results from Table 2 do not allow making a clear decision between VSN or quantile methods. Nevertheless, as shown in R-I plots (Fig. 1), Qt could be more appropriate for normalization of 2-DE volume data, especially with low intensities.

3.2 Validity of simulated data

To validate the simulated data, we compared the distribution of simulated data with real datasets (Fig. 2). Simulated data with four replicates per condition were compared with Pfl DS and simulated data with six replicates were compared with Pmg DS. The distribution of standard deviation and the R-I plot is illustrative of real proteomic datasets.

3.3 Comparisons of FDR modes of estimation

Simulations were used to evaluate the effect of FDR mode of calculation. For each mode, 100 simulations were run (with 10 replicates), and the number of significant differentially expressed proteins for each test and each Fdr was computed. To compare Fdr threshold with actual FDR values, 10 different threshold values were used (from 0.05 to 0.9). The real value of the FDR was then calculated and compared with the threshold used (Fig. 3). Results show a good correlation between the threshold and the different calculated FDR for high-FDR values. However, for lower values, which are more crucial in this kind of analysis (acceptable FDR is generally acknowledged to be 10% in proteomics), the robust FDR method tended to underestimate actual FDR, thus increasing the discovery of proteins falsely declared differentially expressed. Nevertheless, the other methods appeared appropriate in the whole range of values, and are thus potentially more useful for these kind of data.

Furthermore, we evaluated tests and FDR mode by running 100 simulations without differential expression (by skipping step 3 of the data simulation process). For 'robust Fdr', an average of 31.24–4.12 spots (depending on the test used) were declared significant, whereas for the other methods, the average proportion of spots declared differentially expressed was <1 for all tests (from 0.08 to 0.85). This clearly indicates that the robust FDR tends to select undifferentially expressed spots when the null proportion is high.

3.4 Influence of the number of replicates depends on the test and the FDR calculation mode

One of the main issues reported in proteomics studies is the low number of replicates per condition. We have simulated data with 20, 10, 8, 6, 4 and 3 replicates to determine how this affected the different statistical tests results in terms of false positives and false negatives (Fig. 4). For all these tests, the FDR cut-off was set to 0.1. To assess the power (or sensitivity) of the tests, the false non-discovery rate (FnDR) was also calculated as the number of false negatives over the total number of significant proteins [FnDR = FN/(FN + TN)].

As previously observed, the robust FDR method clearly underestimated the false-positive discovery, even for large numbers of replicates. The method from Storey also showed a relatively high number of false positives for small numbers of

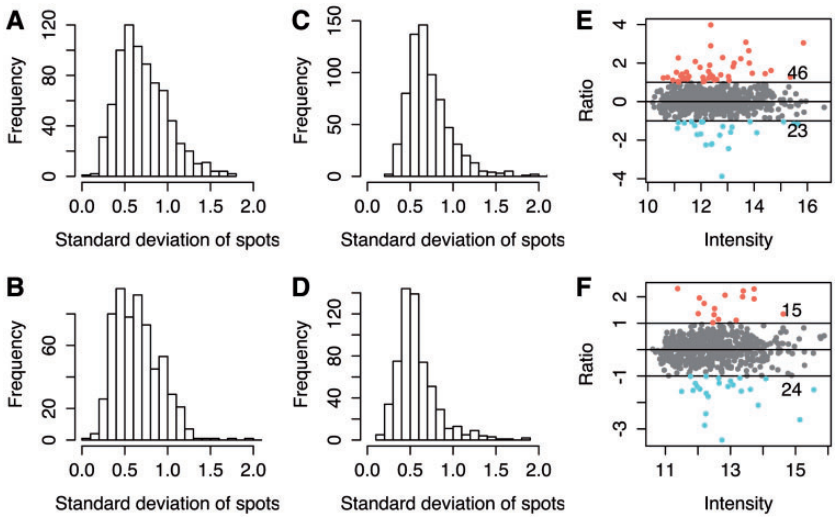

Fig. 2. Comparison between simulated data and real datasets. (**A**) Distribution of standard deviation for Pmg DS. (**B**) Distribution of standard deviation for Pfl DS. (**C**) Distribution of standard deviation for simulated data with 700 spots and six replicates. (**D**) Distribution of standard deviation for simulated data with 700 spots and four replicates. (**E**) R-I plot for simulated data with $n = 6$. (**F**) R-I plot for simulated data with $n = 4$

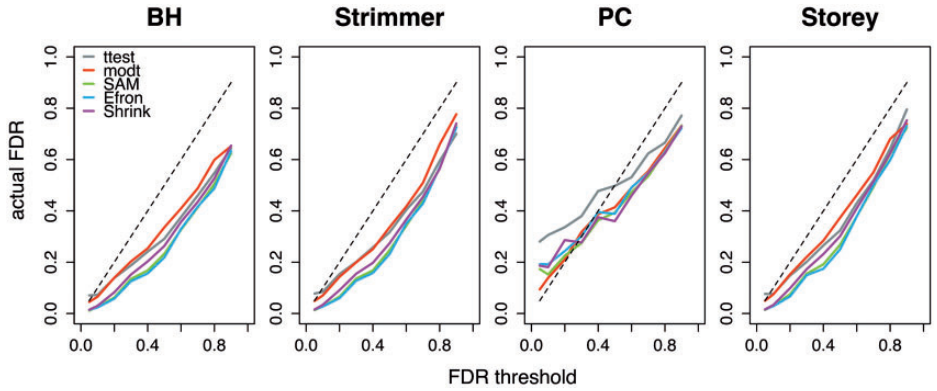

Fig. 3. Threshold versus actual FDR for five statistical tests and four modes of calculation of FDR. Actual FDRs are calculated as the average of 100 simulations with 10 replicates per condition. For comparison purposes, a perfect correlation between threshold and actual FDR is represented as a dashed line. BH (Benjamini and Hochberg, 1995); Strimmer (Strimmer, 2008a); PC (Pounds and Cheng, 2006); Storey (Storey *et al.*, 2004)

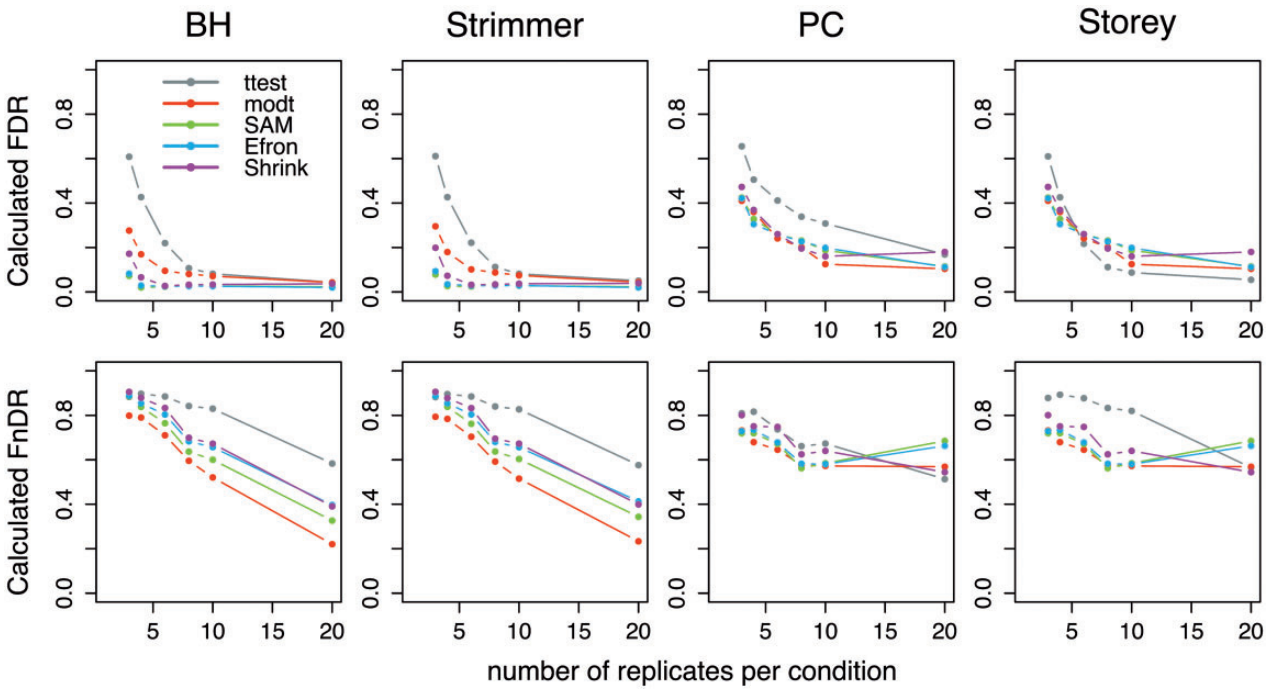

Fig. 4. FDR and False non-Discovery Rate as a function of number of replicates per condition. Calculated values are an average on 100 simulations. BH (Benjamini and Hochberg, 1995); Strimmer (Strimmer, 2008a); PC (Pounds and Cheng, 2006); Storey (Storey *et al.*, 2004)

replicates, but these quickly decrease as the number of replicates increases. We observed similar patterns for classic FDR ('BH') and Strimmer, i.e. the calculated FDR decreased as we increased the number of replicates. Considering the FnDR, we also observed a similar pattern between these two methods, i.e. FnDR decreased as a function of the number of replicates. All the tests displayed this pattern. However, we showed that the classical *t*-test strongly increased the FDR, especially in combination with low numbers of replicates. As for FDR calculation, FnDR did not decrease as rapidly with the classical *t*-test, as compared with the other tests. By contrast, the test that appeared the most sensitive (FnDR systematically lower, regardless of the number of replicates) was the moderate *t*-test. Based on the analysis of both real and simulated data, the moderate *t*-test thus appears to be the most reliable for finding differentially expressed proteins in 2-DE experiments.

4 CONCLUSION

The 2-DE gel-based experiments are often used in comparative proteomics to identify differentially expressed, or accumulated, proteins between two or more conditions. The 2D gels classically allow the visualization of 500–1000 proteins per gel, and replication of experiments is needed to minimize the effects of (i) biological variations in protein expression between individuals and (ii) known technical limits such as differential protein extraction

efficiencies, comigration of proteins, lack of penetration of some proteins in gels or limit of detection of the staining procedure.

Recently, different techniques based both on 2D gel (e.g. DIGE) and gel-free (e.g. iTRAQ) techniques have been developed to reduce these technical biases. However, it seems obvious that, whichever technique is used, statistical analysis remains a crucial step in the proteomics workflow. To date, only a few statistical studies have been dedicated to the specific needs of proteomics, and researchers often use the statistical tools offered in commercial software. The statistical possibilities are restricted, and researchers have to make decisions based only on a *P*- or q-value, without knowing the details of the statistical procedure.

By contrast with proteomics, statistical tools have been widely developed for transcriptomics applications. From a statistician point of view, and notwithstanding different forms of data, the problem is similar in both applications, i.e. extracting statistically significant expression from huge datasets. In this article, we took advantage of many freely available open source statistical tools developed for transcriptomics, evaluated their performance to analyze both real and simulated 2-DE proteomics datasets and developed an R-package adapted to the specific needs of proteomists.

This R-package, called 'prot2D', is freely available as part of Bioconductor (www.bioconductor.org) and includes functions implementing all the methods used in the present article. The Qt method, the FDR calculation method and the moderate *t*-test, that were shown in this article to be the best compromises to analyze proteomics data, are preset in the package with the optimal parameters we determined. For the FDR calculation mode, method from Benjamini and Hochberg and method from Strimmer were shown to be efficient and appropriate for 2-DE volume data.

In all, as statistics is one of the most determining step in the comparative proteomics workflow, and paradoxically, one of the less studied to date, we hope that this new package will help improving future 2-DE-based proteomics studies.

ACKNOWLEDGEMENTS

C. Galland is acknowledged for providing the Pfl dataset. The authors would also like to thank the three anonymous reviewers for their valuable comments and suggestions to improve the quality of the article.

Funding: This research was funded by grants from the Région Bretagne, i.e. the Pemadapt project and a doctoral fellowship to S.A. (Protmar project).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Chang,J. *et al.* (2004) Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. *J. Proteome Res.*, **3**, 1210–1218.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica*, **12**, 111–139.

Efron,B. (2008) Microarrays, empirical bayes and the two-groups model. *Statist. Sci.*, **23**, 1–22.

Efron,B. *et al.* (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.*, **96**, 1151–1160.

Fodor,I.K. *et al.* (2005) Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyderTM. *Bioinformatics*, **21**, 3733–3740.

Galland,C. *et al.* (2013) Comparisons of liver proteomes in the European flounder *Platichthys flesus* from three contrasted estuaries. *J. Sea Res.*, **75**, 135–141.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (**Suppl. 1**), S96–S104.

Karp,N.A. *et al.* (2007) Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. *Mol. Cell Proteomics*, **6**, 1354–1364.

Kim,S. *et al.* (2004) Neuroproteomics: expression profiling of the brain's proteomes in health and disease. *Neurochem. Res.*, **29**, 1317–1331.

Li,F. and Seillier-Moiseiwitsch,F. (2011) RegStatGel: Proteomic software for identifying differentially expressed proteins based on 2D gel images. *Bioinformation*, **6**, 389–390.

Meunier,B. *et al.* (2005) Data analysis methods for detection of differential protein expression in two-dimensional gel electrophoresis. *Anal. Biochem.*, **340**, 226–230.

Miecznikowski,J.C. *et al.* (2011) A comparison of imputation procedures and statistical tests for the analysis of two-dimensional electrophoresis data. *Proteome Sci.*, **9**, 14.

Millioni,R. *et al.* (2012) Operator- and software-related post-experimental variability and source of error in 2-DE analysis. *Amino Acids*, **42**, 1583–1590.

Morris,J.S. (2012) Statistical methods for proteomic biomarker discovery based on feature extraction or functional modeling approaches. *Stat. Interface*, **5**, 117–135.

Morris,J.S. *et al.* (2010) Evaluating the performance of new approaches to spot quantification and differential expression in 2-dimensional gel electrophoresis studies. *J. Proteome Res.*, **9**, 595–604.

Opgen-Rhein,R. and Strimmer,K. (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, **6**, Article 9.

Podrabsky,J.E. and Somero,G.N. (2004) Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. *J. Exp. Biol.*, **207**, 2237–2254.

Pounds,S. and Cheng,C. (2006) Robust estimation of the false discovery rate. *Bioinformatics*, **22**, 1979–1987.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.

R Core Team. (2012) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Stessl,M. *et al.* (2009) Influence of image-analysis software on quantitation of two-dimensional gel electrophoresis data. *Electrophoresis*, **30**, 325–328.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. Ser. B*, **64**, 479–498.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. Ser. B*, **66**, 187–205.

Strimmer,K. (2008a) A unified approach to false discovery rate estimation. *BMC Bioinformatics*, **9**, 303.

Strimmer,K. (2008b) fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, **24**, 1461–1462.

Tomanek,L. *et al.* (2011) Proteomic response to elevated PCO_2 level in eastern oysters, *Crassostrea virginica*: evidence for oxidative stress. *J. Exp. Biol.*, **214** (Pt 11), 1836–1844.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wheelock,A.M. and Buckpitt,A.R. (2005) Software-induced variance in two-dimensional gel electrophoresis image analysis. *Electrophoresis*, **26**, 4508–4520.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.