BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 19 2005, pages 3733–3740
doi:10.1093/bioinformatics/bti612

Gene expression

Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder™

Imola K. Fodor[1,*], David O. Nelson[1], Michelle Alegria-Hartman[2], Kristin Robbins[2], Richard G. Langlois[2], Kenneth W. Turteltaub[2], Todd H. Corzett[2] and Sandra L. McCutchen-Maloney[2]

[1]Computation Directorate and [2]Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA

Received on June 15, 2005; revised on July 5, 2005; accepted on August 2, 2005
Advance Access publication August 9, 2005

ABSTRACT

Motivation: The DeCyder software (GE Healthcare) is the current state-of-the-art commercial product for the analysis of two-dimensional difference gel electrophoresis (2D DIGE) experiments. Analyses complementing DeCyder are suggested by incorporating recent advances from the microarray data analysis literature. A case study on the effect of smallpox vaccination is used to compare the results obtained from DeCyder with the results obtained by applying moderated *t*-tests adjusted for multiple comparisons to DeCyder output data that was additionally normalized.

Results: Application of the more stringent statistical tests applied to the normalized 2D DIGE data decreased the number of potentially differentially expressed proteins from the number obtained from DeCyder and increased the confidence in detecting differential expression in human clinical studies.

Availability: The marray and limma packages used here are available from http://www.bioconductor.org/

Contact: fodor1@llnl.gov

1 INTRODUCTION

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is a technology by which thousands of proteins in a biological sample are separated according to their isoelectric points and molecular weights (O'Farrell, 1975; Görg *et al*., 2000; Lilley *et al*., 2002). In theory, each protein is uniquely determined by its response along the two dimensions of separation. Differences in the proteomes of multiple samples can be studied by comparing the expression profiles of the proteins on the gels. In traditional 2D PAGE, each gel contains one sample which is compared with the samples on different gels, introducing high experimental variability.

Ünlü *et al*. (1997) proposed 2D difference gel electrophoresis (2D DIGE) as a method to overcome gel-to-gel variability inherent to 2D PAGE. More recently, 2D DIGE has been commercialized through the Ettan DIGE System of Amersham Biosciences (now a part of GE Healthcare), thanks to the development of the three size and charge-matched, spectrally resolvable CyDye fluors Cy2, Cy3 and Cy5. Gels using the DIGE method contain three samples labeled with the three distinct fluorescent dyes Cy2, Cy3 and Cy5. Typically, two dyes are used to label two different biological samples of interest. The third dye can be used to label the 'internal standard' which is a pooled mixture of all the samples used in the experiment, and is identical on all gels. The power of the internal standard is in its potential to adjust for the variability between gels and thus make the data across the experiment more comparable. The DeCyder differential analysis software is a part of the Ettan DIGE System, and is used for analyzing the data and quantifying the differential expression of the proteins (Tonge *et al*., 2001; Alban *et al*., 2003; Amersham, 2003).

Although there are fundamental differences in 2D DIGE and gene-expression microarray technologies, many of the difficulties encountered in the analysis of 2D DIGE data are similar to problems that arise in the analysis of microarray experiments: proper normalization of the data within and between the gels (arrays), multiple hypothesis testing and the quest for improved test statistics that exploit the common information across the proteins (genes) (Huber *et al*., 2002, 2003; Smyth *et al*., 2003b; Dudoit and Yang, 2003; Cui and Churchill, 2003). Since data from 2D DIGE experiments exhibit similar characteristics to microarray datasets, we adapted methods developed by researchers in the microarray field to address statistical challenges in analyzing proteomic data from 2D DIGE.

Earlier studies based on DeCyder version 4.0 proposed robust statistical methods and normalization techniques to complement the analytical tools in DeCyder (Kreil *et al*., 2004; Karp *et al*., 2004). We offer additional improvements in the assessment of differential protein expression by combining related normalization methods with novel statistical tests, based on a study with DeCyder version 5.01.

2 APPROACH

To investigate the response of the human proteome on exposure to smallpox vaccination, a proteomic study involving five human subjects, before and at five time points after vaccination, was undertaken. Based on literature indicating the advantages over other 2D gel methods (Tonge *et al*., 2001; Alban *et al*., 2003), 2D DIGE was selected as the technology platform. Blood samples were collected from five volunteers at six time points before and after vaccination, with informed consent under the Institutional Review Board approval from Lawrence Livermore National Laboratory. The samples were prepared and labeled following the manufacturer's

*To whom correspondence should be addressed.

Table 1. 2D DIGE experimental design. Each gel had three samples, two corresponding to a subject sample with time of collection indicated (labeled with Cy3 and Cy5) and a pooled standard that was common on all gels labeled with Cy2

Time	Subject				
	S_1	S_2	S_3	S_4	S_5
T_1: 1 h prior T_2: 1 h post	Gel 1	Gel 4	Gel 7	Gel 10	Gel 13
T_3: Day 1 T_4: Day 3	Gel 2	Gel 5	Gel 8	Gel 11	Gel 14
T_5: Day 7 T_6: Day 14	Gel 3	Gel 6	Gel 9	Gel 12	Gel 15

protocol for 2D DIGE and included the removal of the six proteins with highest abundance (Chromy *et al*., 2004). Details of the sample processing are available from the authors. The resulting 30 samples were arranged on 15 gels as shown in Table 1. The 30 biological samples (five subjects, six time points) were analyzed by 2D DIGE in triplicate, resulting in 45 total gels. In two replicates, on any given gel, the sample corresponding to the earlier sampling time was labeled with Cy3, whereas the sample corresponding to the later time was labeled with Cy5. In one replicate, the dyes were swapped. All gels contained an identical third sample, the pooled standard labeled with Cy2. The scientific goal was to identify proteins that were differentially expressed in response to smallpox vaccination, as a model for smallpox. The aim of the present study was to investigate the results obtained with DeCyder and indicate possible improvements in proteomic data analysis.

DeCyder version 5.01 was used for spot detection and matching across the gels (Amersham, 2003). Both the Differential In-gel Analysis (DIA) and the Biological Variation Analysis (BVA) modules were used: the former to codetect and quantify the spots on a given gel in terms of the ratios of the Cy3 and Cy5 sample volumes to the standard Cy2 volume, and the latter to match the spots and standardize the ratios across the gels accounting for the observed differences in the Cy2 sample volumes on the gels. For each gel, the spot boundaries obtained from the Cy2 image were copied over to the images of the other two samples on the same gel. Since the internal standard was identical on all gels, the software performed the matching only on the internal standard images labeled with Cy2, without introducing sample-to-sample differences into the matching. The master gel was chosen as the gel with the most spots. The other spot maps were matched to the master image with a proprietary 'pattern recognition algorithm that matches one single spot in one gel to a single spot in another gel based on its neighboring spots' (Amersham, 2003). To increase the accuracy of the automatic gel-to-gel matching, careful manual landmarking was performed as recommended in the software documentation.

The volume of a spot for a given dye is defined as the fluorescent intensity of the corresponding dye integrated over the area of a spot. Normalized volume refers to the volume normalized across the three dyes and across the gels. One of the outputs DeCyder provides is the ratio of the normalized volumes, also called the standardized abundances,

$$\begin{cases} R_{pg} = \text{VolCy5}_{pg}/\text{VolCy2}_{pg}, \\ G_{pg} = \text{VolCy3}_{pg}/\text{VolCy2}_{pg}, \end{cases} \quad (1)$$

for each spot p and gel g in the experiment. VolCy5_{pg} represents the normalized volume of spot p on gel g in the Cy5 sample and similarly for the other two dyes.

The statistical analyses in DeCyder are based on the standardized protein log abundances, which are defined as the log10 of the standardized abundances. In theory, the standardized log abundances follow a normal distribution and are comparable across all spots and gels.

The output from DeCyder was exported and analyzed in the R computing environment (http://www.r-project.org/).

2.1 Fitting linear models to assess the differential expression of proteins

The goal of the study was to detect proteins that showed differential expression post-vaccination. Thus, all pairwise comparisons among the six time points were of interest.

DeCyder provides two choices for determining if a protein is differentially expressed between two groups: one based on the fold change and the other on the P-value from the traditional Student's t-test. Fold change is calculated as the ratio of the average standardized abundances corresponding to the two samples. If $\bar{S}_{p1}$ and $\bar{S}_{p2}$ denote the average standardized abundance of protein p in groups $i = 1$ and 2, respectively,

$$\bar{S}_{pi} = \frac{\sum_{R_{pg} \in \text{Group}_i} R_{pg} + \sum_{G_{pg} \in \text{Group}_i} G_{pg}}{|R_{pg} \in \text{Group}_i| + |G_{pg} \in \text{Group}_i|}, \quad (2)$$

then the corresponding fold change is

$$F_p = \begin{cases} +\bar{S}_{p1}/\bar{S}_{p2} & \text{for} \quad \bar{S}_{p1} > \bar{S}_{p2}, \\ -\bar{S}_{p2}/\bar{S}_{p1} & \text{for} \quad \bar{S}_{p1} < \bar{S}_{p2}. \end{cases} \quad (3)$$

A k-fold expression increase/decrease is reflected in a $+k/-k$ value of F_p; no change corresponds to $F_p = 1$.

A common way to assess the differential expression of the proteins is to combine the two measures and find the proteins that exceed a predetermined fold change with a predetermined significance.

In the microarray literature it has been shown that in order to test for the differential expression of many genes in parallel, the traditional Student's t-test can be improved upon (Cui and Churchill, 2003). One common approach is to adjust the gene-specific standard deviation estimates with adjustment factors calculated from a larger set of genes. The idea is to take advantage of the fact that the same model is fit across all genes. The detail lies in specifying how the gene-specific parameters and variances differ. Improved statistics based on empirical methods have been suggested in Baldi and Long (2001) and Efron *et al*. (2001). The moderated t-statistic introduced in Lönstedt and Speed (2002) and further explained in Smyth (2004) (http://www.bepress.com/sagmb/vol3/iss1/art3) is based on a hierarchical, hybrid classical/Bayes model and has been shown to follow a t-distribution under certain assumptions.

In addition to the traditional t-statistics, the moderated t-statistics, as implemented in Smyth *et al*. (2003a), was also used in this study in order to determine the differential expression of proteins. The problem was cast in a general linear modeling framework which facilitated testing using both methods. Consider the model

$$y_{pij} = \alpha_{pi} + \epsilon_{pij}, \quad (4)$$

where y_{pij} is the standardized log abundance of replicate j at time T_i of protein spot p, α_{pi} is the unknown expression level of protein

spot p at time T_i and ϵ_{pij} is a random error, for $p = 1, \ldots, 2384$ (number of spots), $i = 1, \ldots, 6$ (number of time points) and $j = 1, \ldots, 15$ (number of replicates at each time). To follow the analysis with DeCyder, the 3 replicates of the 5 subjects were treated as 15 replicates.

For a given spot p, let $\mathbf{y}_p$ denote the vector of the 90 observations at that spot, ordered according to time: the first 15 values are the replicates at time T_1, followed by the 15 replicates at times T_2, T_3, T_4, T_5 and T_6. Similarly, let ϵ_p denote the corresponding vector of random errors. If $\alpha_p = (\alpha_{p1}, \alpha_{p2}, \ldots, \alpha_{p6})^{\mathrm{T}}$, then the model in Equation (4) can be written in matrix terms as

$$\mathbf{y}_p = \mathbf{X}\,\alpha_p + \epsilon_p, \tag{5}$$

where the design matrix $\mathbf{X}$ has size 90×6, and its i-th column has 15 ones in its $i \times 15$th positions for $i = 1, \ldots, 6$, and is zero everywhere else.

Testing the equality of the expression levels at different times can be easily specified with appropriate contrasts, or linear combinations of the parameters. For example, testing the null hypothesis that the expression level of spot p at time T_1 is equal to the expression level at time T_2,

$$H_0 : \alpha_{p1} = \alpha_{p2}, \tag{6}$$

is equivalent to

$$H_0 : \beta_{p12} = 0, \tag{7}$$

where

$$\beta_{p12} \doteq C^{\mathrm{T}}\,\alpha_p = (-1\ 1\ 0\ 0\ 0\ 0)\,\alpha_p. \tag{8}$$

For each spot in the experiment, the 15 pairwise comparisons among the six time groups were performed, using both the traditional (corresponding to the results from DeCyder) and the moderated t-statistics.

2.2 Normalizing the standardized log abundances

The distribution of the standardized log abundances showed systematic biases within the gels and had different ranges across the gels. Since both of these problems have been encountered by the microarray analysis community, methods developed to address these issues in microarrays were investigated. Specifically, the limma Norm package from the Bioconductor project (Smyth *et al*., 2003a) was used.

To perform the additional normalizations, the standardized abundances in Equation (1) were first transformed into the $M - A$ space, where

$$\begin{cases} M_{pg} = \log_2(R_{pg}/G_{pg}), \\ A_{pg} = 1/2 \log_2(R_{pg} \times G_{pg}). \end{cases} \tag{9}$$

A_{pg} measures the Average, and M_{pg} (Minus) the difference between the intensities of the two samples (samples labeled with Cy3 and Cy5, respectively) on a log scale at spot p on gel g. Assuming that the majority of the proteins were not differentially expressed between the two conditions, the plot of M_{pg} versus A_{pg} (MvA) for a given gel should result in a random scatter around the zero-line with no systematic trends. Observed systematic variations may be the result of different labeling efficiencies for the Cy3 and Cy5 dyes, as well as different scanning settings and gel effects. In microarrays, dye imbalances often vary according to the average spot intensity A (Smyth *et al*., 2003b). The MvA plots for the 45 gels exhibited systematic trends which depended on the value of A (Fig. 4a and 4b); therefore, local intensity-dependent regression lines through the data were fitted using the loessFit function in R. Next, the M-values were replaced by the residuals from the fit which resulted in pattern-free MvA plots (Fig. 4c and 4d). The second normalization step used boxplots for between-gel normalization (Fig. 5). It involved comparing the ranges of the regression-corrected M-values across the 45 gels, and scaling them so that the middle 50% of the data on each gel spanned the same range.

Let $\tilde{M}_{pg}$ and $\tilde{A}_{pg}$ denote the corrected values after the MvA normalization within gels and boxplot normalization between gels. Next, the inverse transformation of Equation (9) was used to transform $\tilde{M}_{pg}$ and $\tilde{A}_{pg}$ back to the original RG scale, and obtain the normalized standardized abundances $\tilde{R}_{pg}$ and $\tilde{G}_{pg}$ corresponding to Equation (1). The standardized abundances from DeCyder were thus further normalized.

The linear model fitting described in Section 2.1 was repeated at each of the spots, using the log10 of $\tilde{R}_{pg}$ and $\tilde{G}_{pg}$ as the response variable in Equation (4). The model was identical to Equation (5), except that the data at each spot consisted of the 90 normalized standardized log abundances instead of the 90 standardized log abundances.

2.3 Adjusting the P-values

Another challenge in the analysis of 2D DIGE data that is shared with the microarray data analysis community is the massive multiple hypothesis problem (Shaffer, 1995). Regardless of the data used and the testing procedure employed, the resulting P-values need to be adjusted because numerous tests are performed simultaneously. The unadjusted P-values that result from the individual t-tests applied separately at each time point pair and at each spot are too optimistic. At the $\alpha = 0.05$ significance level, 1 every 20 tests is expected to result in a P-value less than α just by chance. As the number of tests increases, so does the number of false positives. Several adjustment methods have been proposed. The simplest one is the Bonferroni correction, which multiplies the unadjusted P-values by the total number of tests performed. A less stringent, but more practical approach for the present case is the false discovery rate method of Benjamini and Hochberg (1995). Let R denote the total number of rejected hypotheses, and V the number of falsely rejected hypotheses, out from the total number of simultaneous tests. Then, the realized False Discovery Rate (FDR) is defined as V/R, for $R > 0$, and 0 otherwise. Since V is unobserved, Benjamini and Hochberg (1995) developed a sequential P-value procedure that controls the *expected* value of the FDR, E(FDR), under the assumption that the test statistics are independent. The resulting process controls E(FDR) at the fixed level α for any joint distribution of the P-values. Although the independence assumption is not always satisfied, the FDR method is often used because of its simplicity. Since its results are preferable over the unadjusted P-values, here the FDR procedure in R was used.

3 RESULTS

Figure 1 displays the standardized log abundance data for one protein spot. Assuming that a protein was present in all the samples and that its corresponding spot was found and matched across all 45 gels, there should be 15 values at each time point: three replicates for each of the five subjects. For spot 1186, the third replicate of gel 8

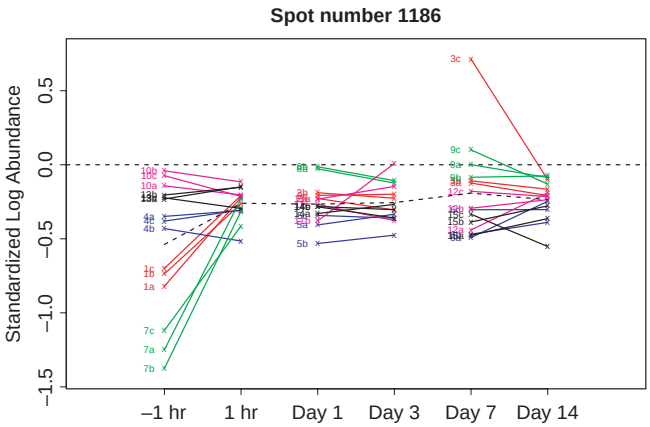

Fig. 1. The standardized log abundance for one spot. Numbers indicate gels, letters stand for replicates, and colors represent subjects. The dotted line connects the averages at the six time points.

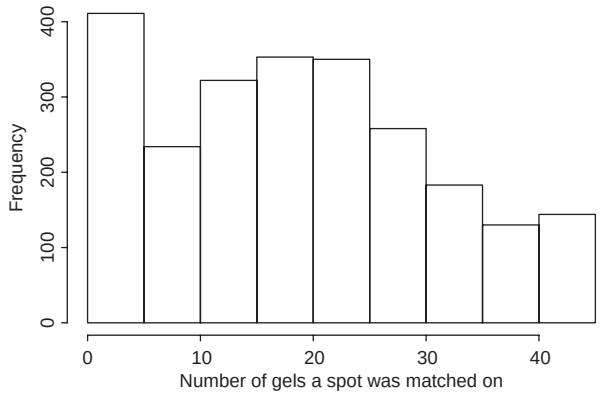

Fig. 2. Histogram of the number of gels a spot was matched on: 2384 spots and 45 gels.

is missing, evidenced by the two green lines connecting Day 1 and Day 3 in Figure 1.

A total of 2384 spots were identified on the master gel, defined to be the gel containing the most spots. Figure 2 presents the histogram of the number of gels a spot was matched on. Fewer than 150 spots were matched on at least 40 of the 45 gels. The less stringent criterion requiring at least five observations at each time point resulted in 1026 spots.

3.1 Results with the Student's *t*-statistic using the standardized log abundances

Table 2 presents the number of spots with >1.5-fold change, and with P-value <0.05, for each of the 15 pairwise comparisons involving the data at two time points. The response was the standardized log abundance and the test was based on the traditional t-statistics. The values in the unadjusted columns used the unadjusted P-values that resulted from performing the traditional t-tests independently at each of the spots and time pairs. The fold changes and the P-values corresponding to the individual spots under the unadjusted heading match the results given by DeCyder. The FDR-adjusted columns refer to P-values that were adjusted for the multiple comparisons. Comparing

Table 2. The number of spots with >1.5-fold change and P-value ≤0.05. Pairwise tests using the standardized log abundances and Student's t-test

	Unadjusted					FDR-adjusted				
	T_2	T_3	T_4	T_5	T_6	T_2	T_3	T_4	T_5	T_6
T_1	7	47	62	53	54	0	8	11	8	11
T_2		47	53	71	59		11	15	8	11
T_3			3	32	49			1	5	13
T_4				55	58				9	19
T_5					8					1

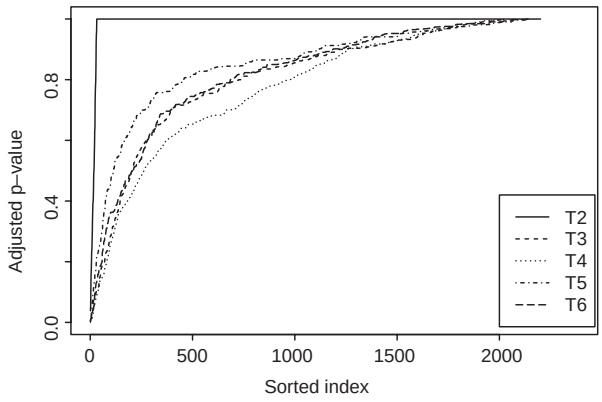

Fig. 3. Sorted FDR-adjusted P-values for the pairwise t-tests that compare the average standardized log abundances at time T_1 to the subsequent time points.

the corresponding numbers under the unadjusted and FDR-adjusted cells in Table 2 illustrates the effect of adjusting for multiple comparisons. The number of 'interesting' spots decreases dramatically after the multiple statistical hypothesis testing problem is addressed.

When aggregating the possibly overlapping results of the 15 pairwise comparisons, a total of 310 unique spots had >1.5-fold change and unadjusted P-value <0.05 in at least one pairwise test. The corresponding number based on the FDR-adjusted P-values was 83.

Figure 3 displays the sorted adjusted P-values from the pairwise t-tests calculated at each spot comparing the five subsequent times to T_1. A possible explanation for the unique shape of the T2 versus T1 curve (solid) compared with the other curves in Figure 3 is the fact that the T2 versus T1 comparisons involved spots from the same gels, whereas the others compared spots from different gels. Example statistics for the number of spots included in the intragel versus intergel comparisons for Subject 1 were: T2 versus T1: 1133 spots (equal to the number of spots on gel 1a that were matched with the spots on the master gel), T3 versus T1: 714 spots (the number of spots on gel 1a that were matched with the spots on both gel 2a and the master gel), T4 versus T1: 714 (same as for T3 versus T1), T5 versus T1: 780 spots (the number of spots on gel 1a that were matched with the spots on both gel 3a and the master gel), T6 versus T1: 780 (same as for T5 versus T1). Similar trends existed for the other subjects as well: more (and better matched spots) for intragel comparisons, fewer (and less well matched) spots for intergel comparisons.

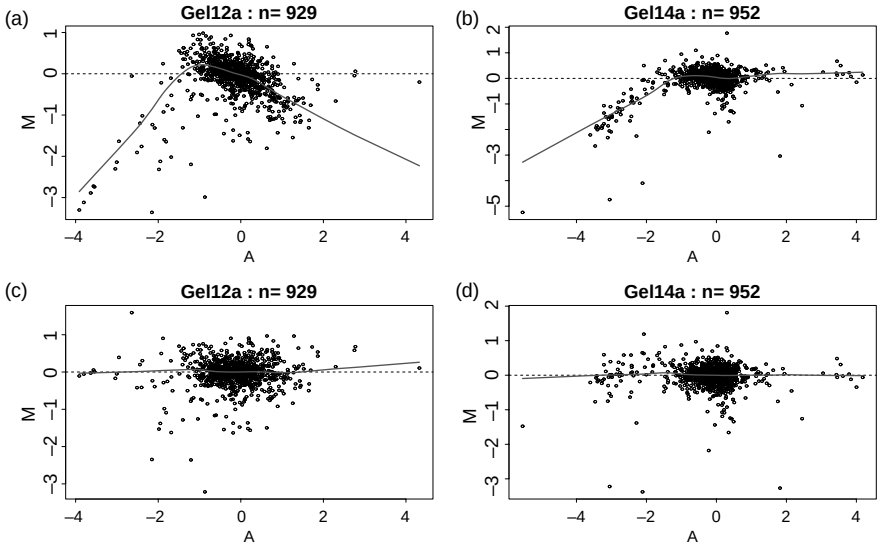

Fig. 4. The MvA plots for gels 12a and 14a: (**a**) and (**b**) based on the standardized log abundances from DeCyder, (**c**) and (**d**) the corresponding results after the loess normalization. The titles reflect the number of spots from the given gel matched to spots on the master gel.

Table 3. The number of spots with >1.5-fold change and FDR-adjusted P-val ≤0.05. Pairwise tests using the moderated t-statistics and (a) the standardized log abundances and (b) the normalized standardized log abundances.

	(a)					(b)				
	T_2	T_3	T_4	T_5	T_6	T_2	T_3	T_4	T_5	T_6
T_1	1	3	4	3	0	1	4	4	0	0
T_2		4	9	5	2		7	7	5	6
T_3			1	2	2			0	4	2
T_4				3	2				4	4
T_5					1					0

3.2 Results with the moderated t-statistic using the standardized log abundances

Panel (a) of Table 3 is similar to the FDR-adjusted panel of Table 2, and presents the corresponding results obtained using the moderated t-statistic along with the standardized log abundances. Results with the unadjusted P-values were generally higher, but overall comparable to the unadjusted results in Table 2. Aggregating the results of the FDR-adjusted P-values from panel (a) of Table 3 from all 15 pairwise tests resulted in 13 unique spots.

3.3 Results with the moderated t-statistic using the normalized standardized log abundances

Figure 4 displays MvA plots for two gels, before (a, b) and after (c, d) the normalizations within the gels. The data for most of the other gels showed similar characteristics. Figure 5 shows the effect of the additional between-gel normalization step. Figure 5a displays the boxplots of the M values based on the output from DeCyder. Differences among the gels are clearly visible, especially for gel 3c which had a higher interquartile range (the middle 50% of the data values within the boxes of the boxplots) than any of the other gels. The unusual distribution for gel 3c was probably caused by problems specific to either that gel or the processing of that gel, as the corresponding distributions for replicates 3a and 3b did not exhibit such anomalies. Figure 5b presents the corresponding results after within-gel normalization. Consequent to the local regression fit, the boxplots in Figure 5b are all centered around zero. However, the interquartile ranges show differences across the gels. The between-gel normalization step brings the interquartile ranges of the gels onto the same scale, as shown in Figure 5c. After the MvA normalization within arrays and boxplot normalization between arrays, the normalized standardized log abundances corresponding to the six time points in the experiment were obtained as described in Section 2.2. Figure 6 displays the result for spot 1186 whose standardized log abundance data were shown in Figure 1.

Panel (b) of Table 3 presents the number of spots with a >1.5-fold change and FDR-adjusted P-value ≤0.05, using the normalized standardized log abundances as the response variable and testing with the moderated t-statistics. Combining the results of the 15 pairwise tests resulted in 13 unique spots. Results with the unadjusted P-values were generally higher, but overall comparable to the unadjusted results in Table 2.

4 DISCUSSION

Figure 7 aggregates the results of the three FDR-adjusted methods in Section 3 in a Venn diagram. The numbers in the circles represent unique spots. Of the eight spots commonly identified by all three adjusted methods, only one spot (2196) had enough observations to be of practical interest from a statistical perspective, loosely defined here as having at least five observations at each time, irrespective of which subject the available replicates belonged to and keeping in mind that subject variability and host response could result in differential expression. Of the three spots commonly identified by TAdjs and NormModTAdj, two (1506 and 1596) contained the required

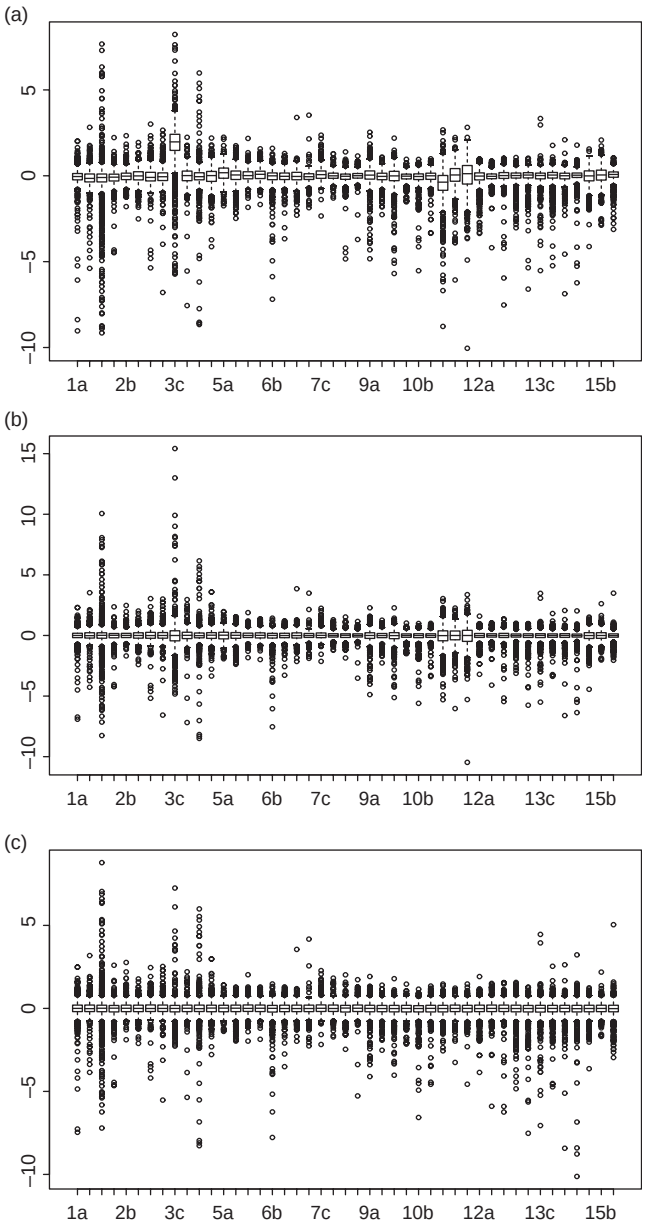

Fig. 5. The boxplots of the M-values for the 45 gels (**a**) before, (**b**) after within-gels and (**c**) after within- and between-gel normalization. The gels are ordered sequentially according to the experimental design in Table 1: the three replicates of gel 1 (1a, 1b, 1c) followed by the three replicates of gels 2 through 15 (15a, 15b, 15c).

number of data points. Being identified by more than one adjusted method suggests a higher confidence that these spots represent proteins that are indeed differentially expressed. Confirmation requires protein identification by mass spectrometry followed by further validation experiments. The three spots identified only by the ModTAdj method, the two spots identified only by the NormModTAdj method and the two spots commonly identified by TAdjs and ModTAdj, each had less than five values per time point, so in this case were not considered although important information may still be found from these patterns.

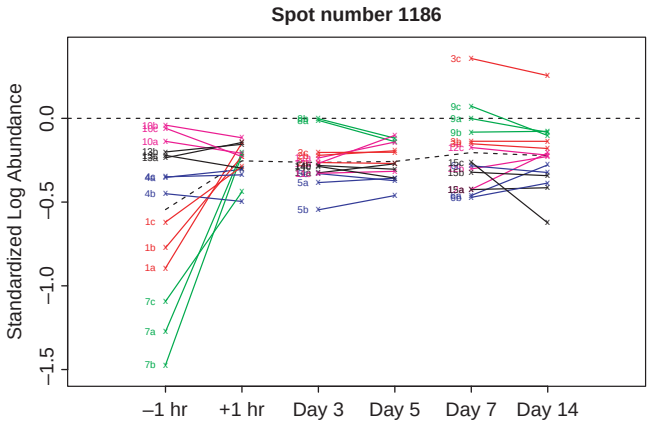

Fig. 6. The normalized standardized log abundance data corresponding to Figure 1.

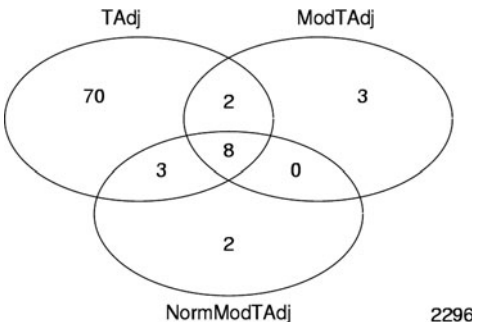

Fig. 7. Venn diagram comparing the results based on the three FDR-adjusted methods in Table 2 (TAdj), Panels (a) (ModTAdj) and (b) (NormModTAdj) of Table 3.

Several factors contributed to the higher complexity of this clinical study, as compared with other published 2D DIGE experiments: (1) the choice of using human blood, one of the most complex proteomes with estimates of 100 000 circulating proteins with a wide dynamic range in concentrations; (2) subject-to-subject variability within the five vaccinees; (3) challenges of variable host immune response; (4) the large number of gels involved. In addition, the gels were prepared in-house. Although the extent of the following challenges is expected to be less severe in simpler experiments, the qualitative conclusions drawn here remain valid for other 2D DIGE studies as well. Our preliminary findings with precast gels (whose reproducibility has been improving in recent years) suggest significant improvements in the quality of the data.

4.1 Normalization

We found evidence for inadequate normalization of the data within and between the gels. Our results agree with other recent findings (Kreil *et al.*, 2004; Karp *et al.*, 2004), and indicate the need to develop better techniques. Since the global characteristics of the data resembled data from microarray experiments, we suggested methods developed in that community as possible ways to improve the normalization of proteomic data from 2D DIGE.

4.2 Accounting for multiple comparisons

Whenever there are multiple hypothesis tests, the observed significance levels have to be adjusted. Here the FDR method was used.

4.3 Matching spots across gels

Although the spot matching rates observed in this study may seem low, there are no reports upon which to compare our results for a human plasma clinical study. Published studies citing 52% (Alban *et al*., 2003) and 67% (Yan *et al*., 2002) of spots matched on gels relied on far fewer gels (12 and 8, respectively) and the use of simpler biological samples (*Escherichia coli*) which would not be affected by genetic variability characteristic to human subjects. In addition, differences in spot matching can be attributed to the wide isoelectric point (pI) and molecular weight (mw) region used in our study: non-linear pI range 3–10, mw range 200–20 kDa. By targeting a narrower pI or mw region, protein spots would be better resolved with improved subsequent matching results. The number of spots specified as an input to the DeCyder algorithm also affects the results. The strategy in this study was to start with a large initial spot number (2500) in order to maximize detection of small-abundance proteins. The large number of spots specified, however, could lead to the inclusion of dust particles or other artifacts. Thus, the current state of the technology is not fully automated, and all potentially interesting spots should be manually verified.

4.4 Spot migration

Microarrays consist of a fixed grid of spots, where each spot contains a unique DNA sequence from a known gene. In contrast, proteins migrate through the gels according to their pI and mw. Genetic differences between subjects and post-translational modifications may result in certain protein spots missing from certain gels, or the 'same' protein migrating slightly differently on the gels. The challenge is to untangle the biological differences in protein expression from differences owing to experimental variation. Spot migration is thus one fundamental difference between microarrays and gels that needs to be addressed, in particular as it relates to spot matching and model development. The mechanistic approach of this paper to ignore spots with poor matching was only a first attempt to understand the data. More sophisticated methods that take into account the underlying biology should be developed, as unmatched spots between subjects may hold information of biological interest.

4.5 Intragel versus intergel comparisons

Although the internal standard is used in 2D DIGE to guarantee that all spots are comparable across all gels, we found evidence to the contrary. The distinct shape of the T2 versus T1 curve, compared with all other time points in Figure 3, points to the different nature of comparing samples from the same gel and comparing samples from different gels. Such differences are most likely because of the imperfect intergel matching. The distinct pattern of the T2 versus T1 curve persisted over the T4 versus T3 and the T6 versus T5 comparisons, but not over the other pairwise comparisons. To minimize the effects of matching, samples of most interest in comparing should be placed on the same gel. Improvements in spot detection and matching should mitigate the differential effects observed in the intergel comparisons. Performing the spot detection separately on each gel image (instead of only on the Cy2 images) may increase the accuracy. The high complexity of the internal standard may have contributed to the poor matching. Perhaps a simpler internal standard consisting of all the T1 samples, or including on all gels an identical T1 reference sample labeled with either Cy3 or Cy5, would have led to superior results. These and other alternatives should be explored, balancing the cost of running the experiment with the quality of the results.

4.6 Statistical modeling

Proper experimental design should be an integral part of any experiment. The design in Table 1 was chosen following recommendations in Amersham (2003). To formulate the optimal design for a given experiment, we advocate interaction with statisticians on the allocation of the samples to the gels, and on proper randomization. Results for microarrays (Kerr and Churchill, 2001) could be extended.

The linear modeling framework of Smyth *et al*. (2003a) used here provides a flexible extension to the simple tests provided in DeCyder. Testing additional hypotheses involving different subsets of the subjects and the time points amounts to specifying different design matrices and contrasts, then proceeding with the estimation as described within. Functionality in R allows one to fit the linear models using robust techniques that minimize the effects of outliers. Accounting for the different number of data points at the different spots is automatically included in the models.

Although the moderated t-test provides an alternative to the Student's t-test for pairwise comparisons, other methods are also possible. From a statistical perspective, a more appropriate way to analyze the data is to fit a mixed effect model at each spot, treating the subjects as five blocks and the gels as two blocks within the subjects (Pinheiro and Bates, 2000). Then, one test at each spot is used to determine if there are any differences among the six time points. Including the block effects improves the estimation of the time effects of interest, and separates the biological replicates from the technical replicates. The two-factor Analysis of Variance (ANOVA) model in DeCyder only supports fixed effects, and is unable to model the random subject and gel effects. Since both the subjects and the gels are samples from larger populations, random effects are appropriate for them. We performed the described mixed-effect modeling at each spot, and found four spots with FDR-adjusted P-value for a time effect <0.05 and at least a 1.5-fold change between any two time points. Of the four spots, one spot (2196) was previously selected by all three adjusted methods. Since spot 2196 was identified by a number of different methods, it has the highest confidence that it is indeed an example of a differentially expressed protein following smallpox vaccination.

The statistical models used here have certain assumptions, such as normality of the errors and independence of the observations. However, these models can be used in an exploratory fashion even if the data exhibit departures from the assumptions (Smyth, 2004). Further model developments should incorporate more realistic assumptions about the data. In addition, they should also take into account the state of the proteins, which will require close collaboration between the proteomics and statistics communities.

5 CONCLUSION

The 2D DIGE technology plays an important role in proteomics, and rigorous data analysis techniques are essential in quantifying the differential expression of proteins between biological samples. Here, we presented readily available statistical methods to improve the analysis of 2D DIGE experiments. Our goal was to offer analytical improvements with small investment to the user. We achieved this

goal by borrowing methods from the microarray literature, and showing their feasibility and suitability to the analysis of 2D gels. To objectively quantify the effects of the proposed techniques, we are currently undertaking a technical variability study using human blood samples.

In addition to the problems shared with microarrays, 2D DIGE presents additional difficulties in spot detection and matching, especially when used in complex studies involving clinical plasma samples. Future advances in image processing and in statistical modeling specific to proteomics will further enhance the quality of 2D DIGE results. Version 6.0 of DeCyder, released after the completion of this study, offers improvements over the version used here in areas such as normalization and adjusting the significance levels in multiple comparisons. We will take full advantage of the latest software in the future.

ACKNOWLEDGEMENTS

We wish to acknowledge our clinical collaborators Harry Lampiris and Lynn Pulliam from the San Francisco Veterans Affairs Medical Center for their assistance with this study. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, with support from the Department of Homeland Security (Biological Countermeasures Program). This work was supported by Laboratory Directed Research and Development funding. UCRL-JRNL-207079.

Conflict of Interest: none declared.

REFERENCES

Alban,A. *et al*. (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, **3**, 36–44.

Amersham (2003) *DeCyder Differential Analysis Software User Manual, Version 5.0*. Amersham Biosciences.

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

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

Chromy,B.A. *et al*. (2004) Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins. *J. Proteome Res.*, **3**, 1120–1127.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA experiments. *Genome Biol.*, **4**, 210.

Dudoit,S. and Yang,Y.H. (2003) Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. *The Analysis of Gene Expression Data: Methods and Software*. Springer, NY, pp. 73–101.

Efron,B. *et al*. (2001) Emprirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Görg,A. *et al*. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053.

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

Huber,W., von Heydebreck,A. and Vingron,M. (2003) Analysis of Microarray Gene Expression Data. *Handbook of Statistical Genetics*, 2nd edn. Wiley, Vol 1, 162–187.

Karp,N.A. *et al*. (2004) Determining a significant change in protein expression with DeCyder during a pair-wise comparison using two-dimensional difference gel electrophoresis. *Proteomics*, **4**, 1421–1432.

Kerr,K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Kreil,D.P. *et al*. (2004) DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics*, **20**, 2026–2034.

Lilley,K.S. *et al*. (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.*, **6**, 46–50.

Lönstedt,I. and Speed,T.P. (2002) Replicated microarray data. *Stat. Sin.*, **12**, 31–46.

O'Farrell,P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021.

Pinheiro,J.C. and Bates,D.M. (2000) Statistics and Computing. *Mixed-Effects Models in S and S-PLUS*. Springer, NY, pp. 8–11.

Shaffer,J.P. (1995) Multiple hypothesis testing. *Ann. Rev. Psych.*, **46**, 561–576.

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

Smyth,G.K., Thorne,N. and Wettenhall,J. (2003a) *LIMMA: Linear Models for Microarray Data User's Guide*. The Walter and Eliza Hall Institute of Medical Reseach.

Smyth,G.K., Yang,Y.H. and Speed,T. (2003b) Statistical Issues in cDNA Microarray Data Analysis. *Methods in Molecular Biology*, Humana Press, Totowa, NJ, Vol. 224, pp. 111–136.

Tonge,R. *et al*. (2001) Validation and development of fluorescence two-dimensional gel electrophoresis proteomics technology. *Proteomics*, **1**, 377–396.

Ünlü,M. *et al*. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

Yan,J.X. *et al*. (2002) Flourescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics*, **2**, 1682–1698.