

Bioinformatics in Gel-based Technologies

Åsa M. Wheelock^{1,2,*}, Craig E. Wheelock¹

¹Karolinska Biomics Center and ²Department of Medicine;

Division of Respiratory Medicine, Karolinska Institutet

ABSTRACT

Fast and robust image processing is a crucial step in quantitative gel-based proteomics. However, although one of the main purposes with a 2DE analysis software is to reduce and correct for experimental and technical variance inherent in the method, the post-electrophoretic analysis itself may also introduce additional variance. As such, the choice of software product or algorithms offered by the chosen software may have a profound effect on the outcome of the study as well as the time spent on computer analysis. In this chapter, we attempt to provide an overview of the main sources of post-electrophoretic variance in current 2DE analysis, as well as tools with which this type of variance can be quantified.

Keywords: proteomics, quantitative 2-dimensional gel electrophoresis, variance, image acquisition, normalization, 2DE analysis software

*Corresponding author: Åsa Wheelock, Ph.D.

Lung Research Lab L4:01, Department of Medicine, Division of Respiratory Medicine
Karolinska Institutet, 171 76 Stockholm, Sweden, Email: asa@para-docs.org

A third of a century ago, a novel method capable of separating 1100 proteins on one single polyacrylamide gel was introduced(22). Although this powerful separation method could detect proteins with abundances as low as 1/10th -1/100th percent of the total protein content in a cell, difficulties associated with visually estimating protein abundances limited its use to primarily qualitative applications. Since 1975, a steady stream of technical improvements has eliminated many of the original limitations(11) of 2-dimensional electrophoresis (2DE) to the point where semi-quantitative studies have become possible. One of the major breakthroughs in this aspect was the development of powerful bioinformatics tools for image analysis and quantification. However, despite the fact that the resulting 2DE analysis software were designed to reduce the experimental variance in order to

enhance the biological variance of interest, the post-experimental software-assisted image analysis has been shown to introduce additional variance into the analysis(32). In this chapter, we attempt to provide an overview of the main sources of post-electrophoretic variance in current 2DE analysis, as well as tools with which this type of variance can be quantified.

BRIEF BIBLIOGRAPHIC REVIEW

The first generation of computational approaches for analysis of 2DE gels started to appear towards the end of the 1970's (Gellab, LIPS, Elsie, TYCHO)(3). These early software were characterized by heavy user interaction and low automation, often requiring a great deal of programming and technical expertise by the user. A decade later, the introduction of graphical interfaces such as windows prompted a second generation of 2DE

analysis software, (Elsie-4, Melanie, QUEST, Gellab-II). However, the original issues of non-user friendly interfaces persisted, rendering this generation of 2DE software as well inaccessible to non-computer scientists(3). It was not until low-cost personal computers equipped with more user-friendly graphical interfaces and powerful processors became widely available in the late 1990's that the third, modern generation of 2DE software were produced(3). Several of these pioneers (Melanie II, CAROL, Z3, and MIR)(3) were subsequently developed into the 2DE analysis software commercially available today, such as ImageMaster 2D Platinum™ / Melanie™, PDQuest™, DeCyder™, Proteomweaver™ and Progenesis™. A major obstacle with this trend away from academic development towards heavy commercialization is that algorithms used in the products become proprietary information, resulting in a black box research approach. However, information regarding the workings of these software can often be inferred from the details of the algorithms in the original academic versions(9). Efforts are still being made to produce custom systems(18), but since their use in existing literature has been very limited, this line of products is beyond the scope of this chapter. Although the order of events may differ, the automated analyses of 2DE gel images utilized in various commercial software programs generally include the following steps: 1) segmentation (spot detection), 2) quantification, 3) background adjustment, 4) image warping, 5) registration (spot matching) and 6) normalization. The challenges in computational 2DE analysis imparted by technical problems in the experimental technique such as artifacts or irregularly shaped or overlapping spots, as well as the purpose, strengths and weaknesses of the algorithms used today are highlighted below.

Spot detection and quantification

Quantification is often the primary goal in proteomics analyses, and accordingly a central point of 2DE analysis software. The segmentation algorithms used in spot detection can be categorized into two classes, parametric and non-parametric. In parametric spot detection, the actual spot shapes in the 2DE map are transformed into the "ideal" spot shape, generally by fitting to Gaussian parameters. The main advantage of this method is that the image can be replaced by a list of spot centers, which greatly reduces the complexity of the data file representing the image. For example, a 2DE image containing 1000 spots is reduced to 28 kilobytes of data, equivalent to a file size of less than 1% of the corresponding pixel-based 2DE map(4). This technique greatly simplifies the subsequent spot matching step, and was therefore used in earlier versions of 2DE analysis software in order to overcome limitations in computational power. Parametric spot detection handles overlapping spots very well(25) and is still used in some products (e.g. PDQuest), although most modern programs utilize non-parametric, pixel-based segmentation algorithms where no constraints on the spot shape are introduced(24). Instead, quantification is performed through a summation of the pixel intensities localized within the defined spot area. The most widely used algorithm for defining the spot boundaries, termed *watershed*, has its origin in the geosciences and the behavior of water in mountain topography. The protein spots in a 2DE gel represents the mountain peaks, and areas where the water pools represent non-protein spots, while the ridges – the watershed – are the protein spots(23).

Background adjustment

The general idea behind background correction is to enhance the signal of the protein stain in the gel image through subtraction of background noise such as non-specific staining or auto-fluorescence of the gel matrix. In order to distinguish

the noise from the signal, spot detection is often performed prior to background subtraction, and non-spot areas are used for calculation of the background noise. Background subtraction can be performed locally or globally across the image. The simplest global strategy is *global constant* background subtraction, consisting of subtraction of all pixels below a set threshold of the maximum intensity. In contrast, global *morphological* algorithms take into consideration all the pixels in the gel that are not part of a detected spot when calculating the background (utilized in e.g. PDQuest and Delta 2D)(1, 32). In local background subtraction algorithms, the pixels in the immediate vicinity of the spot outline are utilized to calculate the background subtraction. Typically the average, lowest or most frequent pixel intensity in the given region is defined as the background level, and is consequently subtracted from each pixel value in the entire spot area (utilized in e.g. Progenesis and ImageMaster). Two-DE gel images are susceptible to the formation of speckles (salt-and-pepper noise), particularly while using post-electrophoretic fluorescent stains. The speckles can be the result of staining of dust particles or SDS residues in- or on the surface of the gel, and some software products offer additional filtering algorithms to reduce this type of noise(4, 9).

Image Warping and Matching

Registration of which protein spot corresponds to the same protein on a different gel image, generally referred to as spot matching, is another central point in 2DE gel analysis. The original spot-based approaches described above, where a list of spot coordinates generated during spot detection was utilized in the alignment process, was complicated by the large gel-to-gel variations in spot migration patterns inherent in 2DE. The protein migration pattern can be distorted by a range of factors, such as an inhomogeneous electric field during

electrophoresis resulting from current leakage, or artifactual modifications of amino acids by constituents of the 2DE separation procedure. Accordingly, spot-based matching strategies (although still used in some programs such as PDQuest) have largely been abandoned and replaced by the use of raw image-based (also called pixel-based) registration strategies, where all the features in the gel image are utilized for matching(9). Additional image warping algorithms, which deform the image in order to counteract geometrical gel-to-gel variations arising from experimental variation, have more or less become standard(1). A common approach, illustrated in Figure 1, is the division of the gel images into a grid system where each cell is stretched to fit a master gel(30). The warping results in both faster and improved accuracy in spot matching. In the past, the warping step has generally been incorporated as an automated algorithm performed in association with the matching step. Recently, a new workflow, perhaps representing a fourth generation of 2DE analysis software, gives the warping algorithm a more central position in the analysis. In these products (e.g. Delta 2D and SameSpots), warping is performed prior to any image analysis, and is said to drastically reduce user time as well as offer improvements in accuracy and reduced subjectivity of the subsequent matching.

Normalization

Despite extensive technical improvements of the 2DE technology over the years, considerable experimental variance still prevails. A large portion of the variance can be ascribed to inefficiencies in protein transfer both during rehydration of the 1st dimension, and from the 1st to the 2nd dimension(35). Variations in protein staining efficiency, both when using covalent tags or post-electrophoretic staining, are other major sources of experimental variance. To counteract the confounding effects upon the ability to quantitatively determine the true biological

variance, various techniques for normalization have been developed and subsequently been incorporated as standard features in 2DE analysis software. One of the first normalization methods utilized in 2DE was *total spot volume (TSV) normalization*(7). In this *global ratiometric* normalization technique, the absolute spot volume is converted into a relative quantity through dividing the individual spot volume of each spot of interest with the sum of all the spot volumes detected in the gel image. A more common variant is *total valid spot volume (VSV) normalization*, where global ratiometric normalization is performed using the sum of a select set of spot volumes (e.g. those validated across all replicate gels). VSV is slightly more robust as it decreases the influence of missing values. Both of these methods are provided in commercially available 2DE analysis software.

In recent years, the development of multiplexing capacity (see DIGE in Chapter 2) has resulted in more sophisticated normalization methods involving *direct ratiometric* normalization. The availability of pI matched, spectrally separated fluors (Cy2, Cy3 and Cy5) allows multiple samples to be co-separated in essentially identical patterns on the same gel, and normalization can be performed through dividing each individual spot volume of one fluor with the corresponding spot volume of a different fluor on the same gel. Most efficient normalization is achieved when one of the CyDyes is used to label an internal standard, ideally created through pooling of all the samples to be analyzed in the study, as this assures that all protein spots are represented in the internal standard, thus facilitating direct ratiometric normalization of all protein spots(2). In addition, the utilization of an internal protein standard allows correction for technical gel-to-gel variability between gels, which significantly improves the statistical aspects of the quantitative analysis. However, the use of a pooled

internal standard requires the sacrifice of up to one-third of the sample. Additional requirements for a pooled standard include that all samples to be included in a given analysis are collected prior to initiating the experiment, something that may prove difficult to fulfill in clinical studies where samples may be collected over an extended period of time. In such circumstances, an alternative internal standard approach similar to that utilized in the ALIS (Alexa-fluor-Labeled Internal protein Standard) methods may be advantageous(33). The ALIS method is a cost-effective alternative to DIGE where a tissue sample resembling the sample of interest – rather than a pooled standard - is fluorescently labeled and utilized as the internal protein standard. Total protein is visualized using a post-electrophoretic fluorescent stain spectrally separated from the ALIS (e.g. SYPRO Ruby), thus facilitating multiplexing. However, as the ALIS approach does not result in super-imposable separation patterns of the internal standard and the sample proteins, a global ratiometric normalization method is used. The main weakness of global ratiometric normalization methods is the lack of ability to correct for local differences in background and separation patterns. Direct ratiometric normalization is dependant on super-imposable separation patterns and thus sensitive to missing values. Two-DE analysis software tend to be better equipped to handle direct ratiometric normalization for multiplexing. Software marketed for multiplexed analysis are also equipped with algorithms capable of correcting for the dual spot migration pattern that is a result of minimal labeling strategies. The use of minimal labeling reduces problems associated with low protein solubility imposed by covalent labeling with the hydrophobic fluorophores used in multiplexing, and implies that only a few percent of the total amount of proteins are labeled with CyDye™. However, one should keep in mind that even if the algorithms can estimate the location of the

“invisible” bulk of the protein during spot picking, problems due to overlapping spot patterns still may arise during subsequent protein identification through mass spectrometry.

METHODOLOGY AND STRATEGY

In recent years, the notion that 2DE analysis software may introduce additional variance into the analysis has been brought to the light(31). Unfortunately, the manufacturers of these products generally do not provide any performance evaluations, and the research community is thus forced to rely on the few evaluations performed by other users (for review, see(32)). Furthermore, no consensus concerning standardized tests for evaluating the overall performance has been reached in the proteomics community, which greatly complicates the issue of comparing different studies and products. The quality of the gel images as well as the particulars of the experimental protocol used, such as the choice of protein visualization method, may have a drastic effect on the performance of the software. Nevertheless, in order to determine the power of study results, it is necessary to evaluate the performance of spot detection, matching and quantification. This chapter suggests methods for evaluating 2DE analysis data both in terms of the quality of automatic spot matching and how to quantify the variance induced by specific 2DE analysis software. The basic protocol for how to perform software-assisted 2DE image analysis is covered elsewhere(17). For the purpose of this chapter, a standard set of images will be employed. These images are freely available for download from www.pulmonomics.net, enabling readers to regenerate the described analysis as well as employ the same images in evaluating other software products not covered in this review.

Image acquisition

Image acquisition is often trivialized in 2DE analysis, even though reproducible image acquisition is as essential as any

other step in minimizing variance. Image acquisition instrumentation can roughly be divided into two categories: i) Via photography, where excitation occurs through illumination with a constant light source (UV or Xenon lamp) followed by detection using a cooled charge-coupled device (CCD) camera, ii) Through scanning, where a laser is utilized for excitation at specific wavelengths (e.g. Ar: $\lambda_{\text{ex}} = 488 \text{ nm}$, ND-YAG: $\lambda_{\text{ex}} = 532 \text{ nm}$, He-Ne: $\lambda_{\text{ex}} = 633 \text{ nm}$) and a photomultiplier tube (PMT) is utilized for detection (for more technical details see(20)). In both types of instrumentation, the use of emission filters is central to reduce background and increase specificity. In the case of CCD-based systems, emission filters may also be used to simulate the specific excitation wavelengths of the laser. Emission filters can either be of long pass (LP) or band pass (BP) type (Figure 2). Long pass filters exclude all light of wavelengths below the given limit, and are generally used to exclude the light source itself. For example, a LP540 filter used in conjunction with an ND-YAG laser prevents light from the laser itself from confounding detection of the emission spectra. The LP filter thus gives a high yield, as the entire emission curve of the fluorophore is collected. However, it also detects the cumulative non-specific emission of the range; hence it does not necessarily result in a higher signal/noise-ratio as the specificity of the signal is reduced. Most biological material autofluoresce around 500-600 nm, and accordingly causes more problems with Ar or NAD-YAG lasers (Cy3, SYPRO Ruby, Deep Purple) than with He-Ne lasers (Cy5, Alexa₆₃₃, Alexa₆₄₇). The non-specific emission can be reduced through the use of a band pass filter. For example, a 560BP30 filter used in conjunction with a ND-YAG laser limits the detection to a narrow window of the emission spectrum centered around 560 nm (Figure 2). BP-filters are a necessity when utilizing multiplexing as it reduces “bleeding” between different flours used on the same gel, given that the

fluors used for multiplexing have sufficient spectral separation to avoid quenching.

Laser scanners have traditionally been ascribed a higher resolution and a broader dynamic range than CCD-based systems. However, the 16-bit format utilized in most modern CCD cameras provides, at least theoretically, a dynamic range of over 4 orders of magnitude, which is typical for most laser scanners. In addition, newer systems (such as Perkin Elmer's ProXPRESS) exhibit 33 μ m resolution(27), which approaches that of laser scanners. The Xenon/UV light source used in this system also makes it compatible with most fluorescent protein stains on the market.(27) The choice between the two types of systems may not have as big an impact on the quality of the image acquisition as it used to, although the higher excitation energy of the laser will likely result in an improved limit of detection.

Regardless of the instrumentation used, the image acquisition step has the potential of introducing variance into the analysis. Unfortunately, no rigorous studies evaluating the actual variance have yet been reported. Ideally, this should be tested for each system through repeated scanning of a single gel, and quantification of the signal with a robust imaging software (e.g. ImageQuant). Furthermore, the PMT setting is bound to have a nonlinear quantitative effect on spot volume and total amount of protein spots detected. This has been investigated in detail for the Typhoon 9400 laser scanner (GE Healthcare, Uppsala, Sweden), showing an exponential relationship between PMT setting and pixel value for all fluors tested (Cy2, Cy3, Cy4, Sypro Ruby) in addition to a constant, gel-specific difference(5). Fluorescent residues from previous scans that adhere to the glass platen, particularly when using post-electrophoretic stains, can also contribute substantially to the variance in background fluorescence. This source of error can be avoided through careful cleaning of the

glass platen with an appropriate solvent (e.g propanol). Remember to check that the solvent is compatible with the material of the glass platen prior to use.

Evaluation of automatic spot detection algorithms

All 2DE analysis software are equipped with some form of automatic spot detection and matching tool. Unfortunately, frequent mistakes in both spot detection and matching are made due to discrepancies in the spot migration pattern, staining artifacts etc. These mistakes are generally corrected through a subsequent manual revision by the user. In addition to being extremely time consuming, this process also introduces a large degree of subjectivity into the results, as the user has to make active choices both in terms of whether a detected spot truly is a spot, and in terms of evaluating correct matching of spots across replicate gel images. For these reasons, it is desirable to optimize the automatic analysis step with the various user-defined settings available in the software. The results can then be evaluated using free-response operator characteristics (FROC) curves. The FROC (or ROC depending on application) curve, a commonly used concept in evaluating the performance of medical diagnosis or pharmaceuticals(26), was introduced as a tool for evaluating 2DE analysis software by Rogers and colleagues(25). The FROC curve represents the relationship between sensitivity and selectivity of the test. In gel-based proteomics, sensitivity is the capability of detecting true spots, while selectivity is the capability of excluding artifacts from being detected as spots. In the standard format, a FROC curve is plotted as the true positive fraction (TPF: (# of correctly detected spots)/(total # of spots in the image) versus FPF (the false positive fraction: (# of artifacts detected as spots)/(total # of artifacts)). However, as it is difficult to calculate the total number of artifacts in a gel image that potentially may be detected as a spot by the algorithms, it is more practical to either

plot the actual number of detected artifacts (FP) in the FROC graph, or to use the highest number of detected artifacts as “total”. In the examples provided in the results section, we have chosen the latter. Regardless of which method you choose, your ideal value will be located in the upper left corner of your graph (Figure 3).

Evaluating software-induced variance

The method presented here is designed to quantify the amount of variance introduced by a 2DE analysis product, and is an adaptation of a previous study by Wheelock et al.(31). All images discussed are available for download from www.pulmonomics.net. Alternatively, new images can be created as described below. Pick one representative 2DE gel image from an analysis set and copy it five times under different file names. Crop the images one by one using an appropriate imaging software (e.g. ImageQuant). It is crucial that the cropping is performed manually with the mouse, as any tool designed for copying the cropping area between images will result in the exact same cropping of the copies. Instead, a gel set with a slight difference in image boundaries should be produced, as this slight shift in boundaries generally results in the software failing to recognize that the replicates in fact are the exact same image. The point of this procedure is to produce “replicates” where both experimental and biological variance has been excluded, in order to reveal any variance caused by the software analysis itself. Next, perform a quantitative analysis of the “identical replicate” gel set. Make sure to test all available background adjustment options, as these may have a significant impact on the quantitative results(31). Review the spot detection and matching manually to assure its accuracy, then export a sufficient set of spot quantities to Excel. Calculate the average spot volume and the coefficient of variance (CV) for each of the matched spots across the five replicates and plot them in a bar graph. By ordering the values according to average spot

volume, you can discover trends in the variance related to spot size. Repeat the analysis with results from different software programs, normalization methods or background subtraction methods. Please keep in mind that while the exclusion of background subtraction can reveal if a certain algorithm introduces bias in the analysis, the omission of background subtraction will result in lower CVs as the total spot volume remains larger for all spots.

The strategy described above relies on a certain amount of sensitivity to the location of the image boundary in the 2DE analysis software. Some programs, e.g. ImageMaster 2D Platinum(32) may be more robust in this matter, which makes it harder to “fool” the software into treating the individual copies of your gel image as replicates. In this case, you can repeat the analysis of a small area of your gel image, and match a subset of spots manually in order to perform your variance analysis.

Analysis of distribution and variance

Most statistical methods used to determine significant alterations are based on the assumption that the data are normally distributed, yet distribution analysis of 2DE data is often neglected. Two types of distribution analyses ought to be performed on each data set: i) the distribution of spot volumes of each individual protein spot across replicate gels (*spot volume distribution*), and ii) the distribution of the resulting variances of the spot volumes across the replicate gels (*variance distribution*), i.e. the values resulting from the “Evaluation of software-induced variance” described above (Figure 4). The distribution pattern can be visually assessed through a histogram (Figure 4, upper panel) or a Q-Q-Normal plot, in which the sample quantiles are plotted against the theoretical quantiles in the corresponding normal distribution (Figure 4, lower panel). A linear correlation implies that the sample is normally distributed, and a formal goodness-of-fit test should be performed for verification.

The Shapiro-Wilk's test was developed for small sample sizes(28), and is a suitable choice for omics experiments (i.e. large-scale data approaches where the number of replicates typically is low (15, 31, 33)). Based on the null hypothesis that the data are normally distributed, the test calculates the correlation of the points in the Q-Q-Normal plot. A rejection of the null hypothesis ($p < 0.05$) thus implies a non-normal distribution, and appropriate transformation of the data should be considered. The most commonly used transformation for 2DE data is log-transform. However, Kreil and coworkers have reported that log-transform may lead to inflated variance at low signal levels(16). The many similarities between mRNA and protein global expression analyses have prompted the exploration of applying transformations common in microarray experiments on 2DE data, and the successful use of arsinh transformation to achieve a normal distribution has been reported by two different groups(12, 15). *Quantile normalization* has also been reported as a successful scaling strategy, particularly for SYPRO Ruby stained gels(7).

Statistical Analysis

A typical 2DE experiment has an identical advantage and disadvantage - a significant amount of data can be generated from a single experiment. Following the collection of these large datasets, it can be challenging to extract the meaningful biological data in a statistically appropriate context. A typical problem is not having sufficient degrees of freedom in the experimental design, in other words, a high number of variables (protein spots) and a small number of measurements (replicate gels)(10, 14). Standard univariate statistical approaches (e.g. Students t-test) could be appropriate if the experiment was designed to only examine alterations in a few proteins. However, univariate test are inherently sensitive to type II error (false positive), and as the investigator usually wants to analyze as

many proteins as possible simultaneously, a multivariate or inductive approach for pattern recognition is often more appropriate. In this case, the experiment is designed as a hypothesis generating approach instead of hypothesis driven. Pattern recognition can be divided into supervised and unsupervised approaches. Unsupervised pattern recognition is useful to determine if data fall into distinct groups where the aim is to detect data similarities, and subsequently no particular biases exist in terms of the group identifications. Examples include cluster analysis consisting of similarity measurement (e.g. correlation coefficients, Euclidean distance and Manhattan distance) linkage (e.g. nearest neighbor and furthest neighbor) and hierarchical clustering (e.g. dendrograms or tree diagrams)(6, 14). Supervised pattern recognition on the other hand attempts to answer a precise question as to the class of an unknown sample and therefore requires a training set of known groupings to construct the model.

A 2DE experiment usually requires an unsupervised approach, of which principal components analysis (PCA) is one of the most common methods. The use of PCA is advantageous as the techniques are designed to transform a large number of possible correlated variables into a smaller number of uncorrelated variables, or principal components (PC). These analyses can be thought of as essentially variable reduction and are used to identify hidden structures in a dataset. Because many 2DE experiments consist of literally thousands of individual proteins, which are often compared over multiple dosing regimens and time-frames, a method capable of dramatically reducing the number of variables to a more manageable data set is advantageous. PCA reduces the dimensionality of the data set through a series of transformations that result in a low-dimensional plot of the data(6). In the analysis, data are structured such that the rows are the samples (in 2DE: the gels) and the columns are the variables (in 2DE:

the protein spots). Many 2DE analysis software packages include multivariate statistical packages, making the analyses straightforward. However, these statistical packages are often rather simplified, and since the spot volume data can be obtained as standardized output from most 2DE software programs into Excel or text format, it may be advantageous to use a specialized statistical package for the analysis.

Output from the PCA analysis includes a series of scores and loadings, in which the relationship of the PCs to the samples is described by the scores and that to the variables is described by the loadings. A PC is a linear function of the original variable that can be thought of as a vector in multidimensional space, with each variable representing an axis(6). The number of significant PCs is ideally equal to the number of significant components. The first PC describes the majority of the variance, the second PC the next greatest portion and so on. The scores have as many rows as the original data matrix and the loadings have as many columns as the original data matrix. The size of each PC is given by the eigenvalue, which can be defined as the sum of squares of the scores. The sum of all nonzero eigenvalues for a data matrix equals the sum of squares of the entire data matrix. The resulting data are often plotted to examine the biological meaning. One of the simplest plots is that of the score of one PC against another. A scores plot can display clustering of distinct groups within the dataset. A loadings plot can then be used to display which loadings (i.e. protein spots) are driving the observed clustering in the scores plot. A biplot superimposes a scores plot and a loadings plot onto a single graph.

All chemometric methods are influenced by the method employed for data preprocessing. An understanding of data preprocessing is essential for correct interpretation of the output from statistical packages. The simplest transformation is of course none at all and the raw data can be employed in the analysis. However,

mean-centering, in which the mean of each variable is subtracted from each variable, is very common. This transformation shifts the scores plot such that it is centered at the origin. However, it can also affect the relative positions of the variables in both the scores and loadings plots. Mean-centering can often reduce the size of the first eigenvalue and influence the apparent number of significant components in a dataset. Another common method for data scaling is standardization, which is performed following mean-centering. Standardizing the data involves dividing each variable by its standard deviation, changing the covariance matrix to the correlation matrix(6). This transformation places all of the variables on approximately the same scale, enabling low values to assume equal significance as high values. If standardization is not performed, then the PCA will be dominated by the most intense (highest abundance) components. There is no general guidance as to whether to use centered or raw data when determining the number of significant components, the most appropriate method being dependent on the nature of the experiment. However, the biological significance of the transformation must be considered when analyzing the data. If the data are standardized, that means that changes in small abundant proteins will contribute significantly to the data analysis. As these proteins are often at the detection limit, they may consist of some artifacts and should be manually confirmed. However, it is often these low abundance proteins that are of interest, explaining why the majority of studies standardize the data.

EXPERIMENTAL RESULTS AND APPLICATIONS

In order to demonstrate how the methods described in the methodology section above are applied to an authentic set of 2DE gels, we have performed an extension of the analyses carried out in a previous study comparing the performance of two 2DE analysis software, PDQuest and

Phoretix 2D Expression (now PG200)(34). In brief, the data set consisted of five technical replicates of an airway epithelial sample separated on pI range 4-7 and a Duracryl gel matrix, and visualized with SYPRO Ruby protein staining as previously described(34). In addition, an “identical replicate” set was generated as described above through copying of one gel image and subsequent individual cropping of the gel using ImageQuant (Nonlinear Dynamics, Sunnyvale, CA, USA). These images were selected for the study based on their high background staining and substantial speckling, as these characteristics challenge most 2DE analysis programs. In the examples below, we analyzed these gel images with the SameSpots/PG240 software (Nonlinear Dynamics, Newcastle, UK).

Evaluation of automatic spot detection algorithms

Following user-guided warping and alignment of the five replicates, automated spot detection was performed through the SameSpots algorithm. A total of 1143 spots were detected and correctly matched across all gels. Manual review of the spots was performed to distinguish the true positives (520 spots) from the false positives (623 spots). All 1143 spots were exported to PG240, and the efficiency of the spot filtering tool in discriminating true spots (TP) from artifacts (FP) was evaluated. Following spot filtering, the true positive fraction (TPF) and false positive fraction (FPF = (1-TNF)) were calculated as described in the methodology section, and plotted in a FROC curve. The results from three different spot filtering criteria at a range of different settings are shown in Figure 3 (spot volume>50,000-250,000; peak height>500-2500; and spot radius>4-12). None of these spot filters resulted in good discrimination of true spots from artifacts, as evidenced by the proximity of the graphs and the cutoff line.

Evaluating software-induced variance

In contrast to the previously performed study using PG200 and PDQuest(34), no manual editing using semi-automatic editing tools was required following spot detection and matching with SameSpots, as all spots were correctly matched. Quantitative evaluation of the software-induced variance as well as the experimental variance was evaluated on 416 manually reviewed true spots using all five background subtraction algorithms (no-background-subtraction, lowest-on-boundary, average-on-boundary, mode-of-non-spot, and progenesis). The resulting CVs for the five identical replicates using progenesis background subtraction are shown as a bar graph in Figure 5 (upper panel). The results for all algorithms are shown as whisker box plots in the lower panel (Figure 5). The complete omission of background subtraction results in the lowest variance. However, inclusion of the background noise in the average spot volume used to calculate the CV results in an overall lower CV than the corresponding variance would cause after background subtraction. Accordingly, a transformation of the no-background-subtraction–results was performed through dividing the spot volumes with the ratio of the no-background-subtraction and lowest-on-boundary spot volumes (Figure 5). The trend remains the same also after transformation, emphasizing the importance of evaluating the choice of background subtraction method as the quantity and nature of the variance in a system has direct implications on the statistical power. Considering that the biological variance within the control group is 40-50%(21), an additional technical variance of 10-20% caused by the analysis software will have a significant effect on the number of replicates needed. Based on a statistical confidence of 80% power and a 0.05 p-value, an overall variance of 50% would require n=3-4, while a variance of 70% would require n=8(21).

The same analysis was also performed for the authentic 2DE gel replicate set (i.e. including technical variance). The software-associated variance accounted for 12-16% of the total variance for all background subtraction methods offered in the SameSpots software, except mode-of-non-spot which caused for 35% of the total variance. These results correspond well to the previous studies performed with Expression (PG200)(34). Similar studies performed on DIGE gel images, the DeCyder software appeared to introduce 30% of the total variance (reported as the unexplained variance)(8), which corresponds well both with previous results from PDQuest(31) as well as the results presented here using the mode-of-non-spot algorithm.

Analysis of distribution and variance

While the reason for the spot volume distribution analysis generally is intuitive, the purpose of the latter may not be as obvious. In quantitative omics analyses, we strive to determine significant biological alterations in expression levels. Due to the sheer number of protein expression levels to be evaluated, it is not always practical, or even appropriate, to perform rigorous statistical testing on each protein spot. Instead we utilize a cutoff level to select proteins of interest. By convention, 1.5-fold or 2-fold changes in expression levels have been utilized as cutoff levels. However, the assumption underlying the concept of the cutoff level strategy is that all of the spots in the data set have a similar behavior in terms of variance and distribution. To determine if a given cutoff level is appropriate we thus have to determine that the variance for individual spots come from the same population, i.e. have a normal distribution. The results from the previous section were analyzed according to the methods described in the Methodology section. It is of great importance to include *all* spots in an evaluation of the overall variance of a system and not just a few “well behaved”

spots. The results from the authentic replicate set (i.e. including technical variance) using lowest-on-boundary background subtraction are displayed in Figure 5. By visual inspection of the bar graph (upper panel), the distribution appears normal. However, formal goodness-of-fit testing revealed that the data set has a significantly non-normal distribution (lower panel). The inserted graph shows the Q-Q plot for the distribution of spot volumes for one of the spots used in the variance distribution analysis. The Shapiro-Wilk test confirms that the spot volume data are normally distributed.

CONCLUSIONS

Historically, the variance induced by computational analysis in gel-based proteomics has been considered neglectable in comparison to the experimental variance. However, the immense technical improvements in recent years have decreased the experimental variance to a point where variance arising from post-electrophoretic analysis becomes prominent, and the portion attributed to the software alone may exceed 30% of the total technical variance(31). Accordingly, the choice of software product or algorithms offered by the chosen software may have a profound effect on the outcome of the study as well as the time spent on computer analysis(18, 31). The lack of a standardized test for the evaluation of software performance makes it difficult to objectively compare the quality of different programs. Towards this end, we suggest that sets of standardized tests in conjunction with standardized sets of gel images be made available to the research community to serve as a benchmark for software and algorithm comparison. The images discussed in the current work are available for download from www.pulmonomics.net. It is our intent that this work spurs the research community and commercial interests to expand on the points raised in this chapter. It is important that quantitative

comparisons of individual software packages are performed in order to evaluate program efficacy as well as enable researchers to quantify and interpret increasingly small variances in biological data sets. We have provided a set of tools for determining the quality of a 2DE image analysis, both in terms of reproducibility and quality of spot detection and matching. These tools may also be utilized to compare the performance of different products, or to optimize the user-defined settings within a program.

FIVE-YEAR VIEW

In spite of the revolution that has occurred in the quality of 2DE-based separation techniques since its introduction 32 years ago, gel-to-gel variability persists in 2DE-gel analysis. As such, fast and robust image processing is a crucial step in quantitative gel-based proteomics, and the fast development in 2DE analysis software in recent years has removed some of the major stumbling blocks in the field. Future efforts are likely to focus on the improvement and evaluation of previously neglected areas such as background subtraction algorithms and statistical analysis as well as decreasing overall analysis time. Furthermore, standardized sets of 2DE gels as well as performance benchmarks for the evaluation of 2DE analysis software need to be established. Attempts have been made to produce sophisticated sets of synthetic gel images that reflect the true characteristics of authentic 2DE gels(24, 25). Rogers and colleagues created a novel model for the creation of synthetic protein spots based on a training set of authentic 2DE spots(24). Unfortunately, no background was introduced to these images, and as evidenced by the results in Figure 5, the background subtraction algorithm can be one of the main sources of variance introduced by the 2DE analysis software(25). Until an ideal set of synthetic gel images reflecting all aspects of the authentic 2DE gel has been constructed, diverse sets of authentic 2DE

gels representative of common distortions (high background, speckles or irregularly shaped spots(31)) ought to be utilized. As user awareness of the effects of varying algorithms upon software performance grows, manufacturers may finally be forced to provide detailed information on the performance of their products. These types of data would greatly assist in software acquisition, enabling potential buyers to evaluate which software is most appropriate for their intended applications. More flexibility and user-defined functions will be on demand as alternative normalization methods are developed. Recent reports that the state-of-the-art protein stain SYPRO Ruby may have polynomial rather than linear correlations to protein quantity may increase the demand for user-defined quantitation curves in commercial 2DE software(33). Furthermore, the limitations of algorithms for background subtraction may be resolved through the introduction of time-resolved fluorescence. The delayed measurement of a fluor's emission excludes auto-fluorescence from plates, reagents or cell debris that generally have very short life-spans (low ns range), and thus eliminate a major source of background noise. In contrast, fluorescent protein stains have very long life-spans (high ns- μ s), and delayed, cumulative detection over time may improve sensitivity(13, 19). The use of both covalently labeled fluors (CyDyes and Alexa-dyes)(13) and ruthenium chelates(13, 19) have been utilized in time-resolved fluorescence applications in related fields. However, the lack of this feature in modern 2DE image acquisition equipment is currently prohibiting the use and development of time-resolved fluorescence in 2DE.

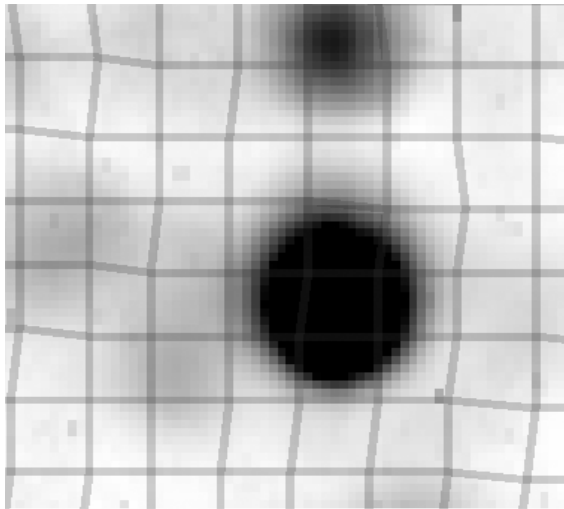
The growth in data acquisition combined with increased quantification capability will continue to expand. These large bodies of data will enable researchers to further understand complex cellular processes and move the field another step closer to comprehension at the organism

level. Proteomics data will enable researchers to model and understand interactions in a cell and to predict the effects of fluctuations upon other intracellular processes. These effects will be seen through, for example, the application of proteomics to web-based models of cellular metabolism such as an electronic cell(29). Proteomics research will assist in the constant march towards true systems biology that will revolutionize personal medicine and fundamentally shift our understanding of biological processes.

REFERENCES

1. Aittokallio T, Salmi J, Nyman TA, Nevalainen OS. 2005. Geometrical distortions in two-dimensional gels: applicable correction methods. *J.Chromatogr.B Analyt. Technol. Biomed. Life Sci.* 815: 25-37
2. Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, 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
3. Appel RD, Palagi PM, Walther D, Vargas JR, Sanchez JC, et al. 1997. Melanie II--a third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. *Electrophoresis* 18: 2724-34
4. Appel RD, Vargas JR, Palagi PM, Walther D, Hochstrasser DF. 1997. Melanie II--a third-generation software package for analysis of 2-dimensional electrophoresis images: II. Algorithms. *Electrophoresis* 18: 2735-48
5. Back P, Bengtsson S, James P. 2005. Automated PreScan Function for Scanning Fluorescently Stained 2D-Gels. *J.Proteome.Res.* 4: 1511-5
6. Brereton RG. 2003. *Chemometrics: Data analysis for the laboratory and chemical plant.* West Sussex, England: John Wiley and Sons
7. Chang J, Van Remmen H, Ward WF, Regnier FE, Richardson A, Cornell J. 2004. Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. *J.Proteome.Res.* 2004.Nov.-Dec.;3(6.):1210.-8. 3: 1210-8
8. Corzett TH, Fodor IK, Choi MW, Walsworth VL, Chromy BA, et al. 2006. Statistical Analysis of the Experimental Variation in the Proteomic Characterization of Human Plasma by 2-Dimensional Difference Gel Electrophoresis. pp. 2611-9
9. Dowsey AW, Dunn MJ, Yang GZ. 2003. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics.* 3: 1567-96
- 10.Engkilde K, Jacobsen S, Sondergaard I. 2007. Multivariate data analysis of proteome data. *Methods Mol Biol* 355: 195-210
11. Gorg A, Weiss W, Dunn MJ. 2004. Current 2-dimensional electrophoresis technology for proteomics. *Proteomics.* 4: 3665-85
12. Gustafsson JS, Ceasar R, Glasbey CA, Blomberg A, Rudemo M. 2004. Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *Proteomics.* 4: 3791-9
13. Handl HL, Gillies RJ. 2005. Lanthanide-based luminescent assays for ligand-receptor interactions. *Life Sci.* 77: 361-71
14. Karp NA, Griffin JL, Lilley KS. 2005. Application of partial least squares discriminant analysis to 2-dimensional difference gel studies in expression proteomics. *Proteomics.* 5: 81-90
15. Karp NA, Lilley KS. 2005. Maximising sensitivity for detecting changes in protein expression: Experimental design using minimal CyDyes. *Proteomics.* 5: 3105-15
16. Kreil DP, Karp NA, Lilley KS. 2004. DNA microarray normalization methods can remove bias from differential protein expression analysis

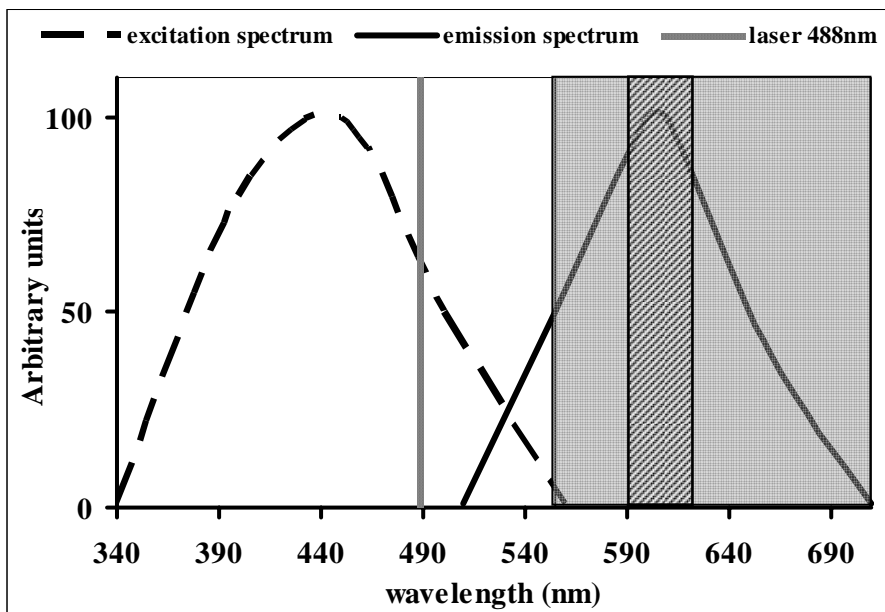
- of 2D difference gel electrophoresis results. *Bioinformatics*. 20: 2026-34
17. Levanen B, Wheelock AM. In press. Ptoubleshooting Image analysis. In *Two-dimensional Electrophoresis Protocols*, ed. D Sheehan, R Tyther. Totowa, NJ: Humana Press
 18. Marengo E, Robotti E, Antonucci F, Cecconi D, Campostrini N, Righetti PG. 2005. Numerical approaches for quantitative analysis of two-dimensional maps: a review of commercial software and home-made systems. *Proteomics*. 5: 654-66
 19. McKie A, Vyse A, Maple C. 2002. Novel methods for the detection of microbial antibodies in oral fluid. *Lancet Infect.Dis*. 2: 18-24
 20. Miura K. 2001. Imaging and detection technologies for image analysis in electrophoresis. *Electrophoresis* 22: 801-13
 21. Molloy MP, Brzezinski EE, Hang J, McDowell MT, VanBogelen RA. 2003. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics*. 3: 1912-9
 22. O'Farrell PH. 1975. High resolution two-dimensional electrophoresis of proteins. *J.Biol.Chem* 250: 4007-21
 23. Pleissner KP, Hoffmann F, Kriegel K, Wenk C, Wegner S, et al. 1999. New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases. *Electrophoresis* 20: 755-65
 24. Rogers M, Graham J, Tonge RP. 2003. Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images. *Proteomics*. 3: 887-96
 25. Rogers M, Graham J, Tonge RP. 2003. Using statistical image models for objective evaluation of spot detection in two-dimensional gels. *Proteomics*. 3: 879-86
 26. Scheipers U, Perrey C, Siebers S, Hansen C, Ermert H. 2005. A tutorial on the use of ROC analysis for computer-aided diagnostic systems. *Ultrason Imaging* 27: 181-98
 27. Scrivener E, Boghigian BA, Golenko E, Bogdanova A, Jackson P, et al. 2005. Performance validation of an improved Xenon-arc lamp-based CCD camera system for multispectral imaging in proteomics. *Proteomics*.: In Press
 28. Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality. *Biometrika*: 591
 29. Tomita M. 2001. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol* 19: 205-10
 30. Veesser S, Dunn MJ, Yang GZ. 2001. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*. 1: 856-70
 31. Wheelock AM, Buckpitt AR. 2005. Software-induced variance in two-dimensional gel electrophoresis image analysis. *Electrophoresis* 26: 4508-20
 32. Wheelock AM, Goto S. 2006. Effects of post-electrophoretic analysis on variance in gel-based proteomics. *Expert Rev.Proteomics*. 3: 129-42
 33. Wheelock AM, Morin D, Bartosiewicz M, Buckpitt AR. 2006. Use of a fluorescent internal protein standard to achieve quantitative two-dimensional gel electrophoresis. *Proteomics*; .
 34. Wheelock AM, Morin D, Goto S, Buckpitt AR. 2005. Sources of variance in gel-based quantitative proteomics studies. Proc. SIPZOO: From Genome to Proteome in Animal Science. 40:74-90
 35. Zhou S, Bailey MJ, Dunn MJ, Preedy VR, Emery PW. 2005. A quantitative investigation into the losses of proteins at different stages of a two-dimensional gel electrophoresis procedure. *Proteomics*. 5: 2739-47



1
2
3
4
5
6
7
8
9

Figure 1

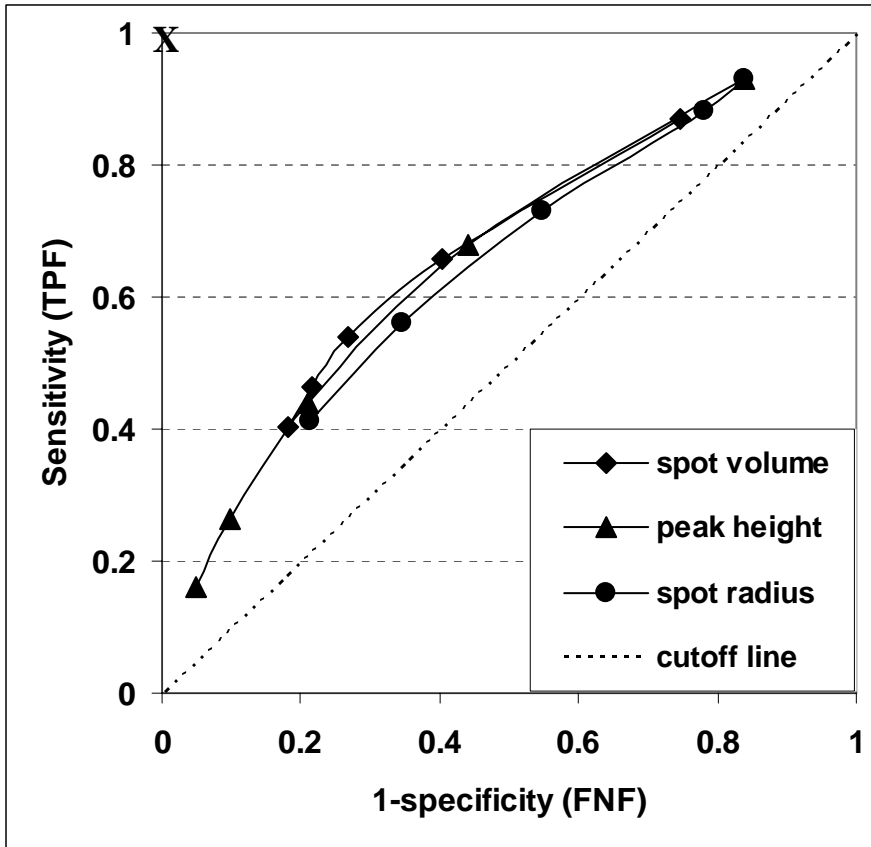
Display of the grid model used in many warping algorithms in order to counteract distortions in separation pattern that cause significant gel-to-gel variations even in replicate gels run under identical conditions.



10
11
12
13
14
15
16
17

Figure 2

Illustration of the use of long-pass (LP) and band-pass (BP) emission filter, in conjunction with SYPRO Ruby protein staining. The dotted line represents the excitation spectra, while the solid line represents the emission spectra. The use of a LP filter (here 555nm) allows cumulative collection of all emission above a given wavelength (entire grey area), resulting in high sensitivity. The use of a BP filter (here 610BP30) selectively collects the emission from a narrow range, here 30nm centered around 610nm (striped area), resulting in high specificity.



1

2

Figure 3

3

4

5

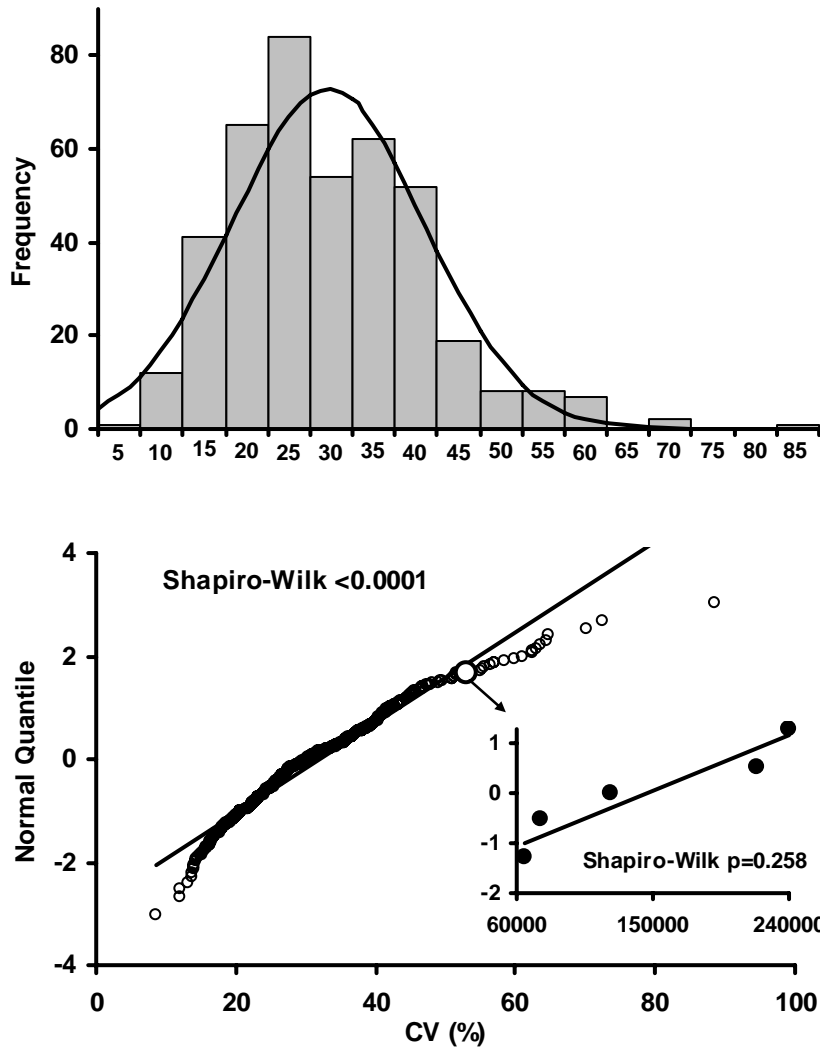
6

7

8

9

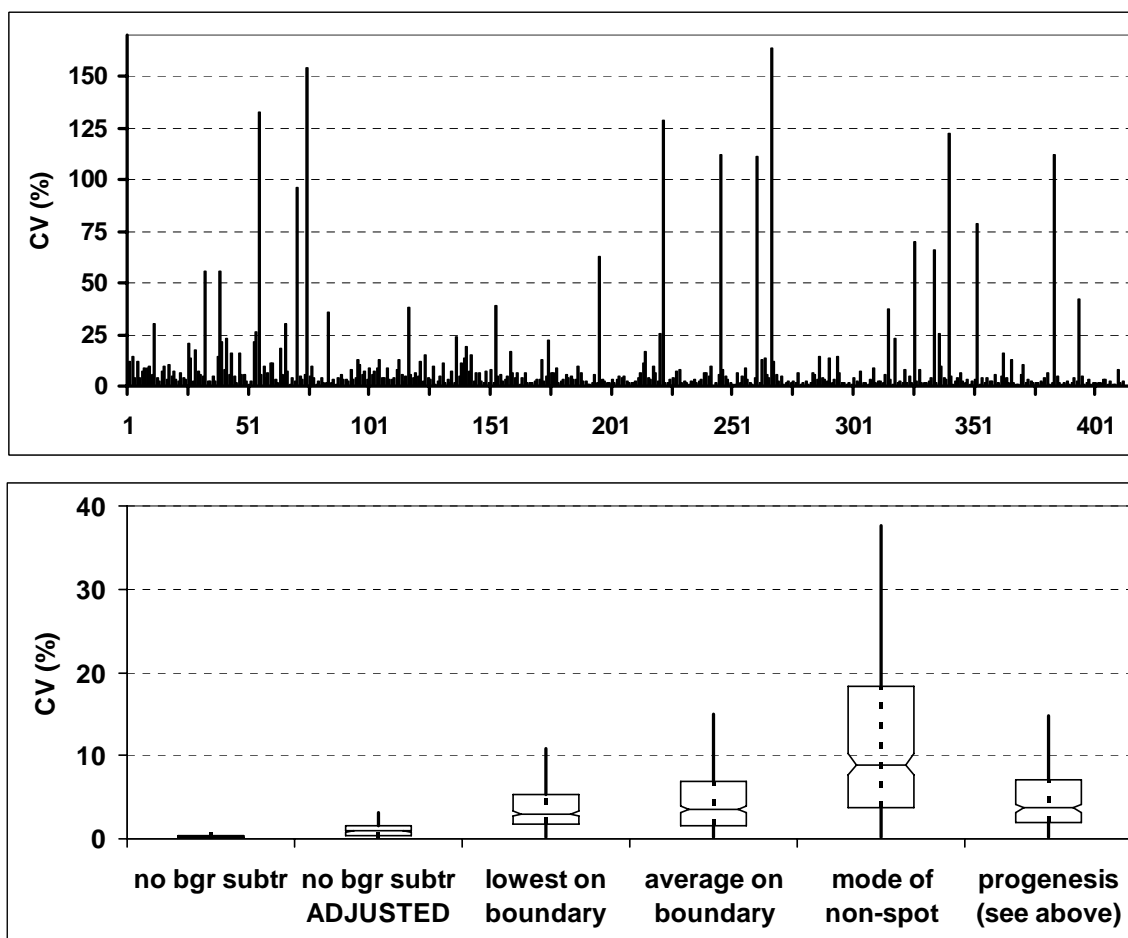
FROC curves were generated in order to evaluate the various filtering options available in Progenesis 240 following analysis with the SameSpots software. A range of values for minimum spot volume, peak height and spot radius were used. The diagonal line represents no discrimination, and the “X” mark in the upper left corner represents the ideal result achieved through manual spot editing. None of the filters were particularly effective in filtering out artifacts (false positives) from true spots (true positives), although spot volume appeared to offer a slightly better selection criteria.



1
2
3
4
5
6
7
8
9
10

Figure 4

The bar graph in the top panel shows the distribution of the variance in spot volumes resulting from an analysis of five replicate 2DE gel images (i.e. including technical variance). The overlaid curve in black shows the corresponding normal distribution. The main graph in the bottom panel shows the Q-Q-Normal plot for the same data set. Formal goodness of fit test using Shapiro-Wilk W test revealed that the data set has a significantly non-normal distribution. The inserted graph shows the Q-Q-Normal plot for the distribution of spot volumes for one of the spots used in the variance distribution analysis. The Shapiro-Wilk Wtest confirms that the spot volume data are normally distributed.



1

2 **Figure 5**

3 The top panel shows the variance in spot volumes across five copies of the same gel image
 4 following analysis with SameSpots, using the Progenesis background subtraction algorithm.
 5 Each bar represents the CV of one protein spot across the five identical replicates, and the
 6 bars are ordered according to increasing average spot volume. The bottom panel shows a box
 7 plot representing the CVs for the same experiment for each of the available background
 8 subtraction algorithms (lowest-on-boundary, average-on-boundary, mode-of-non-spot and
 9 progenesis). In addition, the results from complete exclusion of background subtraction (no
 10 bgr subtr), as well as an adjustment of these results to correct for the decrease in CV caused
 11 by the larger spot volumes resulting from exclusion of background subtraction. Each box
 12 contains the interquartile range with the mean marked inside, and the whiskers showing the
 13 range of CVs following exclusion of outliers.