Clinical Chemistry 49:10
1615–1623 (2003)

Proteomics and
Protein Markers

Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization

Kevin R. Coombes,[1*] Herbert A. Fritsche, Jr.,[2] Charlotte Clarke,[3] Jeng-neng Chen,[4] Keith A. Baggerly,[1] Jeffrey S. Morris,[1] Lian-chun Xiao,[1] Mien-Chie Hung,[4] and Henry M. Kuerer[5]

Background: **Recently, researchers have been using mass spectroscopy to study cancer. For use of proteomics spectra in a clinical setting, stringent quality-control procedures will be needed.**

Methods: **We pooled samples of nipple aspirate fluid from healthy breasts and breasts with cancer to prepare a control sample. Aliquots of the control sample were used on two spots on each of three IMAC ProteinChip® arrays (Ciphergen Biosystems, Inc.) on 4 successive days to generate 24 SELDI spectra. In 36 subsequent experiments, the control sample was applied to two spots of each ProteinChip array, and the resulting spectra were analyzed to determine how closely they agreed with the original 24 spectra.**

Results: **We describe novel algorithms that (*a*) locate peaks in unprocessed proteomics spectra and (*b*) iteratively combine peak detection with baseline correction. These algorithms detected ~200 peaks per spectrum, 68 of which are detected in all 24 original spectra. The peaks were highly correlated across samples. Moreover, we could explain 80% of the variance, using only six principal components. Using a criterion that rejects a chip if the Mahalanobis distance from both control spectra to the center of the six-dimensional principal component space exceeds the 95% confidence limit threshold, we rejected 5 of the 36 chips.**

Conclusions: **Mahalanobis distance in principal component space provides a method for assessing the reproducibility of proteomics spectra that is robust, effective, easily computed, and statistically sound.**

Recently, the scientific community began using proteomics to improve our understanding of the biological processes that underlie cancer. To cite one example, *The Lancet* published an article in February 2002 on the use of proteomics spectra to distinguish serum from healthy individuals from the serum of women with ovarian cancer *(1)*. The authors of that study are planning a large-scale clinical trial to determine the practical utility of this technology as a diagnostic tool. Several of those authors have stressed the importance of their quality-control (QC)[6] steps. If proteomics spectra are to be applied in a clinical setting, the collection and processing of the data will need to be subject to stringent QC procedures. Because of the complicated nature of the data, however, even preclinical studies can be expected to benefit from a rigorous QC process.

Breast cancer, in spite of the widespread use of screening mammography, remains the most prevalent carcinoma in women; it causes the death of more than 45 000 women in the US every year. Analysis of the biochemical and cellular contents of breast ductal fluid has recently gained attention because it offers a noninvasive method to study the local microenvironment associated with the development and progression of breast carcinoma *(2–5)*. The potential benefits have led to renewed interest in nipple aspiration of breast ductal fluid (NAF), a method introduced in the 1970s. In nipple aspiration, a simple, handheld suction cup is placed on the nipple and used to obtain a concentrated fluid fraction of breast secretions quickly and noninvasively. Our group recently began studying a new application of nipple aspiration. By ob-

Departments of [1] Biostatistics, [2] Laboratory Medicine, [4] Molecular and Cellular Oncology, and [5] Surgical Oncology, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd., Box 447, Houston TX 77030.

[3] Ciphergen Biosystems, Inc., 6611 Dumbarton Circle, Fremont, CA 94555.

*Author for correspondence. E-mail krc@odin.mdacc.tmc.edu.

Received December 11, 2002; accepted July 11, 2003.

[6] Nonstandard abbreviations: QC, quality control; NAF, nipple aspirate fluid; 2D-PAGE; two-dimensional polyacrylamide gel electrophoresis; SELDI, surface-enhanced laser desorption and ionization; SPF, simple peak finding; SPDBC, simultaneous peak detection and baseline correction; KS, Kolmogorov–Smirnov; and PCA, principal components analysis.

taining ductal fluid samples from a breast containing a known carcinoma and the same patient's healthy contralateral breast, we can compare the protein expression profiles of these samples. We believe that nipple aspiration may provide a practical method for identifying clinically relevant tumor markers that may be useful in risk stratification, diagnosis, treatment monitoring, and detection of cancer recurrence. Using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), we have previously demonstrated that ductal fluids contain large amounts of protein *(6)*. Systematic comparisons of breast ductal fluid samples obtained by nipple aspiration from women with unilateral breast cancer revealed significant differences in ductal fluid protein expression between the breast with cancer and the breast without cancer in each patient. 2D-PAGE revealed that breast ductal fluid contains more than 1000 separate protein species. This finding suggested that ductal fluids from breast cancer patients might be useful for high-throughput biomarker discovery. Because the breasts are a paired organ system, we might discover significant differences by conducting systematic comparisons of their NAF when cancer develops in one breast.

Surface-enhanced laser desorption and ionization (SELDI) uses a chemically modified surface on an array that binds a subset of proteins to categorize proteomic patterns *(7, 8)*. Significantly, SELDI analysis provides measurements of proteomic patterns in a wide range that includes low molecular masses (<20 kDa). By contrast, 2D-PAGE typically assays proteins in a much higher range of molecular mass. Because SELDI may identify smaller molecules and potentially relevant protein metabolites, conventional proteomic techniques and SELDI analysis of NAF may represent complementary methods of biomarker discovery.

To use a new technology such as SELDI for biomarker discovery, strict QC methodologies must be developed. In this report, we describe a method that we have developed for monitoring the quality of a series of proteomics experiments that use SELDI technology. SELDI experiments are performed on ProteinChip® arrays (Ciphergen Biosystems, Inc.); each array contains eight identical spots that are used to process samples in parallel. In our method, we first run a series of replicate experiments to establish a reproducible proteomics profile for a control sample. In subsequent experiments, we reuse the same sample on two spots of each array as an internal QC for that array. We have developed a robust biostatistical method that allows us to quantify how closely a newly acquired proteomics spectrum matches the profile derived from the initial set of experiments. Our method depends on the measured heights of many peaks in the spectra; it therefore incorporates quantitative information about heights of peaks in addition to qualitative information about their existence. These measurements, in turn, depend on the processing algorithm used to identify and quantify peaks. We have also developed improved algorithms, which are described here, to process raw proteomics spectra to identify and quantify peaks. The improvements in peak detection and quantification contribute directly to our ability to assess the quality of proteomics spectra.

Materials and Methods

SAMPLE COLLECTION AND PREPARATION

Patients. Patients who presented to the Nellie B. Connally Breast Center at the University of Texas M. D. Anderson Cancer Center were eligible to participate in this prospective investigation (as approved by the Institutional Review Board) if they had biopsy-confirmed unilateral primary invasive breast cancer and gave written consent to undergo bilateral nipple aspiration. Patients were excluded from participation if they had previously undergone subareolar surgery that might have disrupted the terminal ductal system.

Ductal fluid collection. Ductal fluid was collected by nipple aspiration with a handheld suction cup similar to nonpowered breast pumps used to express milk from lactating women. This simple device consists of a plastic cup connected to a section of polymer tubing. The tubing is attached to a standard syringe that is used to create a gentle vacuum. This device was originally used and described by Sartorius et al. *(9)* and was purchased for this study from Product Health, Inc.

Before aspiration was attempted, the nipple was cleansed with a small amount of Omniprep paste (D.O. Weaver and Co.) to remove any keratin plugs and then cleansed with an alcohol pad. A small amount of lotion was placed on the breast, and the breast was gently massaged from the chest wall toward the nipple for 1 min. The suction cup was then placed over the nipple, and the plunger of the syringe was withdrawn to the 5–10 mL mark until ductal fluid was visualized. Fluid droplets were collected in a 10-μL graduated micropipette (Drummond Scientific Co.). NAF samples were obtained from both breasts, and the presence of NAF and the NAF volumes obtained were recorded for each patient and each breast. Immediately after collection, the NAF samples were rinsed into centrifuge tubes containing sterile phosphate-buffered saline supplemented with the protease inhibitors 4-(2-aminoethyl)-benzenesulfonylfluoride HCl (AEBSF; 0.2 mmol/L), leupeptin (50 mg/L), aprotinin (2 mg/L), and dithiothreitol (0.5 mmol/L). The total protein concentration was measured with the RC DC assay (Bio-Rad Laboratories).

GENERATION OF SELDI SPECTRA

For analysis on ProteinChip arrays, each sample was first normalized to 2 g/L with 50 mmol/L Tris-HCl, pH 7.0, and then diluted 1:1 with a binding buffer (50 mmol/L Tris-HCl, pH 7.0; 330 mL/L isopropanol; 170 mL/L acetonitrile; 0.5 mol/L NaCl), vortex-mixed, and incubated on ice for 20 min.

Before use, the IMAC3 arrays were loaded with copper

by incubating 5 μL of a 100 mmol/L copper sulfate solution on each spot for 20 min in a humid chamber. Each IMAC-Cu^{2+} array was then washed with doubly distilled H_2O. Finally, each spot was incubated with 5 μL of 100 mmol/L sodium acetate, pH 4, buffer for 5 min and washed again with water. Before samples were added, each spot was prewetted with 2 μL of binding buffer for 5 min, and the buffer was removed. We incubated 2 μL of each sample on the arrays for 30 min in a humid chamber followed by three 5-min washes with 5 μL of Tris-HCl, pH 7.0, containing 1 mL/L Triton X-100. After the final buffer wash, each spot was quickly washed twice with doubly distilled H_2O and allowed to air dry. Lastly, we applied 0.5 μL of sinapinic acid (Ciphergen) dissolved in 200 μL of 500 mL/L acetonitrile–05 g/L trifluoroacetic acid twice on each spot and allowed the spots to air dry.

Molecules retained on the surfaces were visualized by reading the spots of each array in a ProteinChip System time-of-flight mass spectrometer (PBSII; Ciphergen). Each array was read under conditions suitable for low mass (up to 30 kDa) and again for higher mass acquisition (30–150 kDa) to scan the spot, and 100 shots were averaged with automatic data collection protocols in the Peaks (Ciphergen) software program. Low mass spectra were calibrated on a mass calibration curve constructed with bovine insulin (5733.6 Da), bovine cytochrome *c* (12 230.9 Da), and equine myoglobin (16 951.5 Da), and the higher mass spectra were calibrated on a mass calibration curve constructed with horseradish peroxidase (43 240 Da), bovine serum albumin (66 430 Da), and chicken conalbumin (77 490 Da).

DESIGN OF QC EXPERIMENTS

To generate a QC sample for all future SELDI experiments containing noncancer- and cancer-associated ductal fluid proteins and metabolites, aliquots of NAF from healthy breasts and breasts with cancer were pooled. The QC sample was divided into aliquots at the start of the experiment and stored at −80 °C. A large number of ProteinChip arrays was purchased from the same manufacturing lot. Three IMAC ProteinChip arrays, each containing eight spots, were randomly selected from this lot and used for experiments on 4 successive days. On each day, an aliquot of the QC sample was incubated on two previously unused spots on each array, and proteomics spectra were generated. All subsequent experiments used the QC sample on spots A and B of an array from the same lot, along with six different experimental samples on the remaining spots (C–H). All samples were prepared by the same technician and hand-spotted on the arrays.

ALGORITHMS

Simple peak finding (SPF). For the SPF algorithm, the input is a spectrum consisting of mass equivalents and intensities at a sequence of equally spaced time points. Inputs (and outputs) are treated as double-precision floating-point numbers. The output is a list containing the time locations of potential peaks and their associated left-hand and right-hand bases. The parameters are time resolution (T) and mass resolution (M). The algorithm is as follows:

1. Using first differences between successive time points, locate all local maxima and local minima in the spectrum.
2. Using the median absolute value of the first differences to define noise, eliminate all local maxima whose distance to the nearest local minimum is less than the noise.
3. Combine local maxima that are separated by fewer than T clock ticks (or time intervals; the default value is T = 3) or by less than M relative mass units (default value of M is 0.05% of the smaller mass). Retain the highest local maximum when combining nearby peaks.
4. Refine the locations of the nearest local minima to both the left and right of each of local maximum. Because of flat regions, the right-hand minimum for one peak may be at a different location than the left-hand minimum of the next peak.
5. Compute the slopes from the left-hand local minimum up to the local maximum and from the local maximum down to the right-hand local minimum. Eliminate peaks where both slopes are less than half the value of the noise.

Simultaneous peak detection and baseline correction (SPDBC). For the SPDBC algorithm, the input is a spectrum consisting of mass equivalents and intensities at a sequence of equally spaced time points. The inputs (and outputs) are treated as double-precision floating-point numbers. The output is a baseline-corrected spectrum and a list of peak locations. The parameters are baseline window width (B), noise window width (N), and signal-to-noise threshold (S). The algorithm is as follows:

1. Use SPF to obtain a preliminary list of peaks and their bases.
2. Remove the peaks and interpolate across the base linearly.
3. Compute the baseline as the local minimum in a window of width B (default value of B = 256 clock ticks).
4. Construct a revised spectrum by subtracting the baseline from the original spectrum.
5. Repeat steps 1–4 using the revised spectrum as input to obtain a new spectrum, which will be returned as the final baseline-corrected spectrum.
6. Run SPF on the baseline-corrected spectrum to obtain a list of candidate peaks.
7. Compute the noise as the median absolute deviation (MAD) from the median in a window of width N around each peak (default value of n = 512 clock ticks).
8. Eliminate peaks where the signal-to-noise ratio is less than some threshold S (default value of S = 3).

HARDWARE AND SOFTWARE

All computations were performed on a Dell Workstation PWS340 with a 1.8 GHz Pentium 4 processor and 1 GB of RAM. Peak finding and baseline correction were imple-

mented in MATLAB (The MathWorks Inc.), Ver. 5.3. Additional statistical analysis was performed in S-Plus 2000 for Windows (Insightful Corp.). All scripts have been tested with more recent versions of MATLAB or S-Plus on both Windows and UNIX machines. Scripts are available from the first author on request.

Results

In one typical low-mass spectrum derived from a sample of NAF and containing ~18 000 time points, the first step of SPF identified ~3500 local maxima. By applying the preliminary noise filter, we reduced the number of local maxima to ~1500. Combining nearby peaks further reduced the number of peaks to slightly fewer than 500. Finally, removing peaks with small slopes reduced the number to 425.

Although SPF is adequate for identifying peak locations, it cannot be used in isolation to compute peak intensities because it does not correct the baseline. For applications that attempt to distinguish between sample types (healthy vs cancer, for example) based on proteomics spectra derived from serum, we believe that peak intensities are important. Thus, in addition to locating the peaks, we must estimate the baseline and remove it from the spectrum. To accomplish this goal, we use the SPDBC algorithm. We applied SPDBC to the typical sample spectrum described above. As mentioned, the first pass through SPF generated a list of 425 peaks. After the first baseline-correction pass, we visually inspected the corrected spectrum. We found that although most of the baseline had been successfully removed, an overall downward trend remained in the middle portion of the spectrum. Consequently, we instituted a second baseline-correction step, which removed this trend. Both the second and third passes through SPF generated lists of 421 peaks. Only 209 of these peaks exceeded a signal-to-noise threshold of 3.

We applied SPDBC to 24 different spectra produced from separate preparations of the same original sample material. The algorithm identified 206–252 peaks in each spectrum. The mean and the median number of peaks identified per spectrum were both 227.5. The peaks identified by SPDBC were not the same in all spectra. To match peaks from one spectrum to another, we pooled the list of detected peaks and again combined peaks that differed in location by 3 clock ticks or by 0.05% of the mass. Using this criterion, we identified a total of 702 distinct peaks that exceeded a signal-to-noise threshold of 3 in at least 1 of the 24 spectra. To obtain consistent peak heights, we normalized each baseline-corrected spectrum by dividing by the total ion current beyond the 2000-Da range.

Our principal goal was to use these 24 spectra to develop a reproducible proteomics profile that would allow us to use this sample as an internal QC for future proteomics experiments. Seven hundred peaks should provide a robust measure of reproducibility, but only if they represent actual peaks that provide independent evidence. To assess independence, we computed the Pearson correlation between the peaks across the 24 spectra (Fig. 1). We found a high degree of correlation (or anticorrelation) between peaks. Nearby peaks, in particular, were frequently highly correlated. Consequently, we looked for ways to reduce the number of peaks that were required to assess the reproducibility of the proteomics spectra.

We first determined the number of times each peak was detected; the results are summarized in a histogram (Fig. S1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol49/issue10/). Sixty-eight peaks were detected in all 24 spectra. At the other extreme, 245 peaks were only identified once. We chose to retain only peaks that were detected at least twice. We next restricted the region where we searched for peaks to filter out noise regions. The raw spectra produced by the Ciphergen SELDI instrument contain intensity values on a scale from 0 to 100. At low mass values (under ~700 Da per charge) where the baseline is large, the intensities frequently saturate, which limits our ability to make accurate measurements of the amount of protein present. Among the 702 peaks, 127 peaks occurred at masses <700 Da. Combining these two conditions, we found a total of 356 peaks that were detected in at least two different spectra in the region above 700 Da per charge. Further analysis used the intensities and locations of these 356 peaks.

We computed the mean and SD of the height of each peak across the 24 replicate experiments (data not shown). Unsurprisingly, we found that peaks with larger mean intensities tended to have a larger SD. One of the characteristics of proteomics spectra, however, is that peaks centered at higher mass values tend to be both broader and lower than peaks centered at low mass values. To

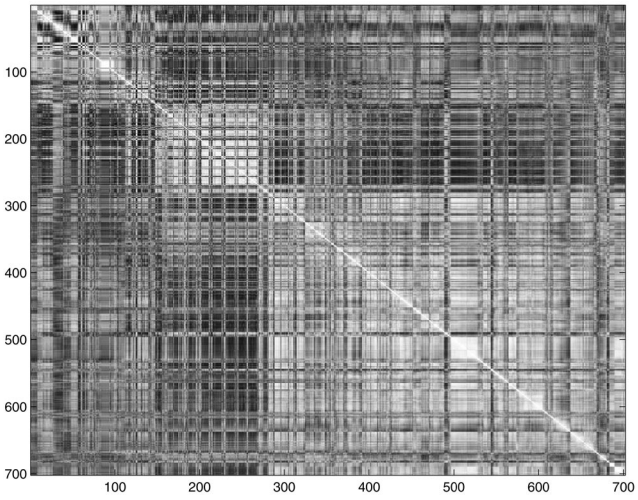

Fig. 1. Correlation matrix of 702 peak heights across 24 replicate spectra.

The intensity scale runs from −1 (*black*) to 1 (*white*). In general, the peaks are highly correlated.

balance the weight given to those peaks in our analysis, we searched for a transformation that would stabilize the variance independently of the mean peak height. We tried a logarithmic transformation, which would have the added benefit in studies that compared different sample types of allowing us to interpret the data directly in terms of fold changes. Unfortunately, zero is a perfectly plausible estimate of the height of an absent peak. To avoid computing the nonexistent logarithm of zero, we explored transformations that replaced the intensity X by a shifted logarithm, $\log(a + X)$, for some choice of the shift constant a. For our data set, these transformations were sensitive to the choice of the parameter a. In addition, each shifted logarithmic transformation distorted the variance in a nontrivial way. We also investigated several different power transformations. To assess whether the transformations produced data that were approximately gaussian in distribution, we prepared quantile-quantile plots (Fig. S2 in the online Data Supplement). We also performed Kolmogorov–Smirnov (KS) goodness-of-fit tests to compare the data after each transformation with the gaussian distribution. In this setting, smaller test statistics and larger P values are associated with data that more closely fit a gaussian distribution. On the original scale, the test statistic was KS = 0.164 ($P < 10^{-120}$). After log transformation, KS = 0.072 ($P < 10^{-118}$), after square root transformation, KS = 0.053 ($P < 10^{-61}$), and after cube root transformation, KS = 0.030 ($P < 10^{-18}$). We concluded that the cube root transformation was most successful (among those examined) at removing the dependence of the variance on the mean peak height (Fig. S3 in the online Data Supplement). After we applied a cube root transformation, the heights of the 356 peaks across the 24 samples remained highly correlated (data not shown).

ANOVA

The initial set of 24 experiments on the QC sample was performed using three IMAC chips on 4 successive days. On each day, the QC sample was placed on two previously unused spots on each chip. This design allowed us to perform a classic ANOVA to determine whether spot-to-spot, chip-to-chip, or day-to-day variation was more important to the reproducibility of SELDI proteomics spectra. A preliminary graphical investigation suggested that day-to-day variation might be more important, with spectra run on the second day being slightly brighter than spectra run on other days (Fig. 2). We then analyzed the data, using a series of linear mixed models. We started with models of the form:

$$Y_{ijkl} = \mu + \mathrm{P}_i + C_j + D_k + (CD)_{jk} + S_{jkl} + \epsilon_{ijkl}$$

where Y_{ijkl} is the observed peak intensity; P_i ($i = 1, \ldots, 356$) is the peak effect; C_j ($j = 1, \ldots, 3$) is the chip effect; D_k ($k = 1, \ldots, 4$) is the day effect; $(CD)_{jk}$ is an interaction term; S_{jkl} ($\mathrm{l} = 1, 2$) is the effect of the two replicate spots for each combination of chip and day; and the residual errors, ϵ_{ijkl}, are assumed to be normally distributed with mean zero and common standard deviation σ. In all models, C, D, CD, and S are treated as random effects, normally distributed with mean zero. Models were fit using the lme function of S-Plus, and the amount of variation attributable to each random effect was assessed using restricted maximum likelihood (REML) estimators of the corresponding variance components.

We obtained similar results from models estimated on three different scales (raw, log-transformed, and cube-root-transformed); we report only the results after cube-root transformation. We ran separate models treating P as either a fixed effect or a random effect. When P was treated as fixed, the residuals contained 90.9% of the variance. When P was treated as random, the peak-to-peak differences explained 91.3% of the variance, with residuals accounting for an additional 7.8%. Thus, in both models, peak-to-peak differences accounted for most of the variability. In addition, the interaction term CD was negligible; therefore, the remaining variation was partitioned between chip-to-chip (6%), day-to-day (26%), and spot-to-spot (68%) differences.

PRINCIPAL COMPONENTS

To provide QC, we needed a summary statistic to measure how far an individual spectrum deviated from the typical spectrum produced from this sample. After much trial and error, we rejected methods that selected a very small number of individual peaks because of the lack of robustness inherent in such a procedure. However, any procedure that uses all 356 peaks must account for the facts, already noted, that the heights of different peaks are highly correlated and that the heights of individual peaks are highly variable. Principal components analysis (PCA) seemed to be ideal in this setting because the goal of PCA is to find linearly independent combinations of the variables (in our case, the heights of different peaks) that explain the maximum amount of variation present in the data.

We performed a PCA on the cube roots of the heights of 356 peaks across the original 24 samples. We found that the first principal component explained 45.7% of the variance, the first six principal components explained 81.2% of the variance, and the first 11 principal components explained 90.4% of the variance (Fig. S4 in the online Data Supplement). Plotting the samples against the first two principal components confirmed the earlier indications that the day effect is larger than the chip effect (Fig. 3).

QC

To assess the quality of our proteomics experiments, we used two spots of every eight-spot ProteinChip to run additional replicates of the same QC sample that was used to derive the original profile. To date, we have run 36 additional chips, so we have performed an additional 72 experiments using the QC sample. Each spectrum was baseline-corrected using SPDBC and normalized to the

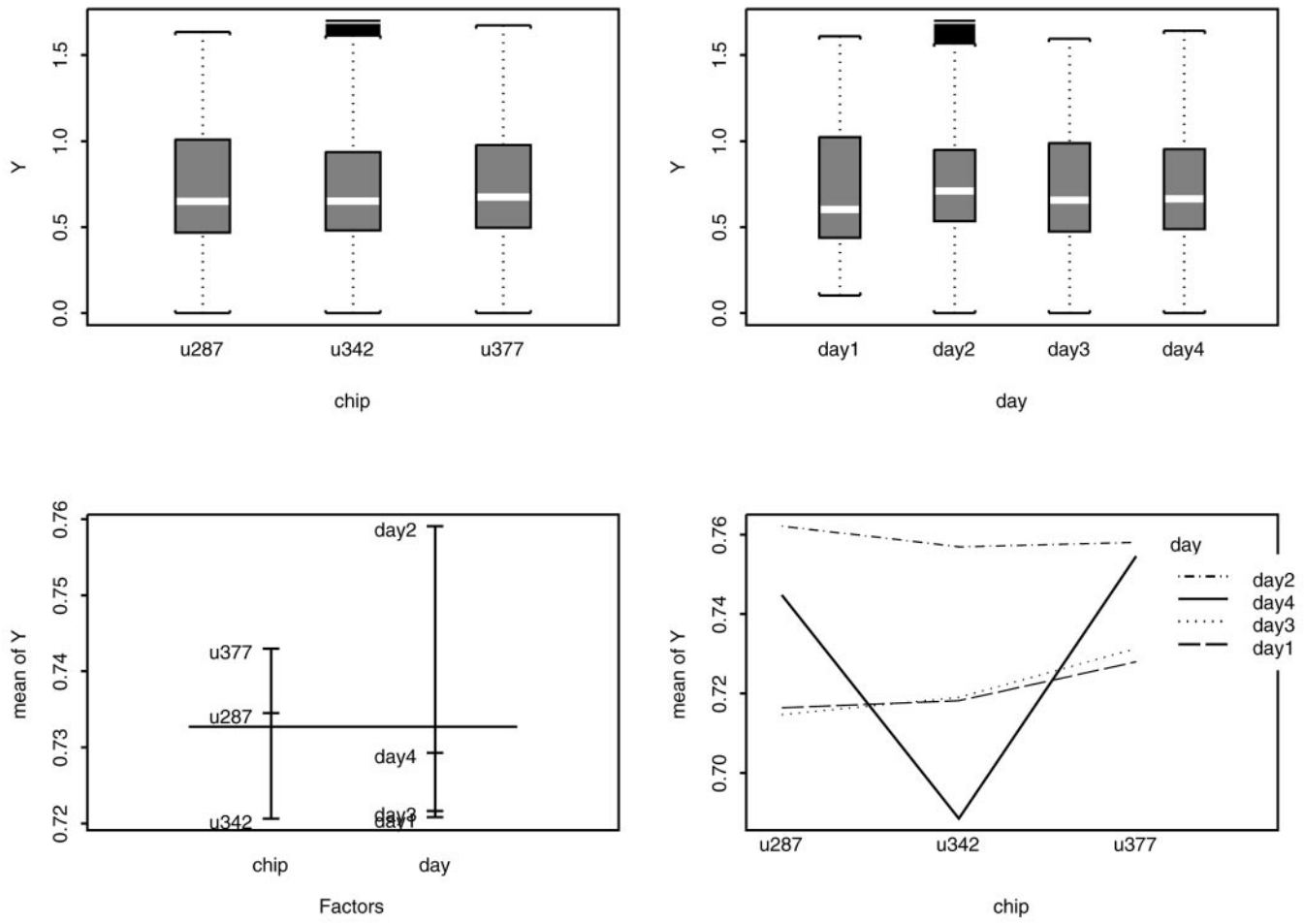

Fig. 2. Relative importance of different chips and different days on the cube roots of the heights of 356 peaks.

The box-and-whisker plots (*top*) show the quartiles of peak heights by chip (*top left*) and by day (*top right)*. (*Bottom left plot)*, mean peak height by chip and day; (*bottom right plot)*, interactions between the factors.

total ion current >2000 Da according to the procedure described above. Heights were determined at the 356 peaks selected from the original 24 spectra and were cube-root-transformed. Each new sample was then projected into the space defined by the first n principal components from the original 24 samples (Fig. 4). Under standard assumptions of normality, the Mahalanobis distance from a control spot to the center of the principal

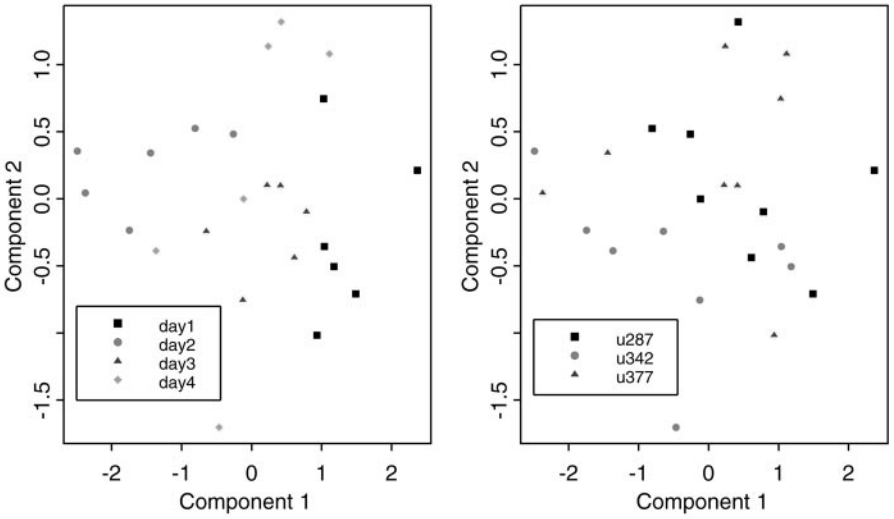

Fig. 3. Spectra plotted against the first two principal components, which explain 60% of the total variance, derived from 356 peaks.

The two plots are identical except for labels. The *left plot* indicates the day on which the experiment was performed; the *right plot* indicates the chip. There is a moderate day effect and a slight chip effect.

component space follows a χ^2 distribution with n degrees of freedom. We decided to reject a chip if the Mahalanobis distance from both control spectra to the center of the six-dimensional principal component space exceeded the 95% confidence limit threshold. Using this procedure, we rejected 5 of the 36 chips.

We visually inspected the QC spectra that were rejected by our method. In every case, they showed substantial deviation from the average of the 24 original QC spectra (data not shown). We also examined spectra from the experimental samples that were analyzed on the rejected chips. In general, these spectra contained fewer peaks than similar samples on chips that were accepted. To see how these differences would affect the downstream analysis, we performed a hierarchical cluster analysis on the experimental samples from four ProteinChip arrays (Fig. 5). Samples from six patients were analyzed on these four chips. Each patient contributed two independent samples; each sample was analyzed on two different chips. Thus, the selected chips included both biological and technologic replicates. Two QC spots and six experimental samples were analyzed on each chip. One chip (C3) was rejected by our QC procedure; the other three were accepted. Significantly, all samples on the rejected chip clustered by themselves on a separate branch of the dendrogram. One should note, however, that the replicate samples within the rejected chip still clustered together, suggesting that some of the biological structure has been retained. By contrast, all biological and technologic replicates clustered together for the three patients whose samples were analyzed on chips that passed our QC procedure. We concluded that our QC procedure successfully identified chips on which technologic artifacts dominated the relevant biology.

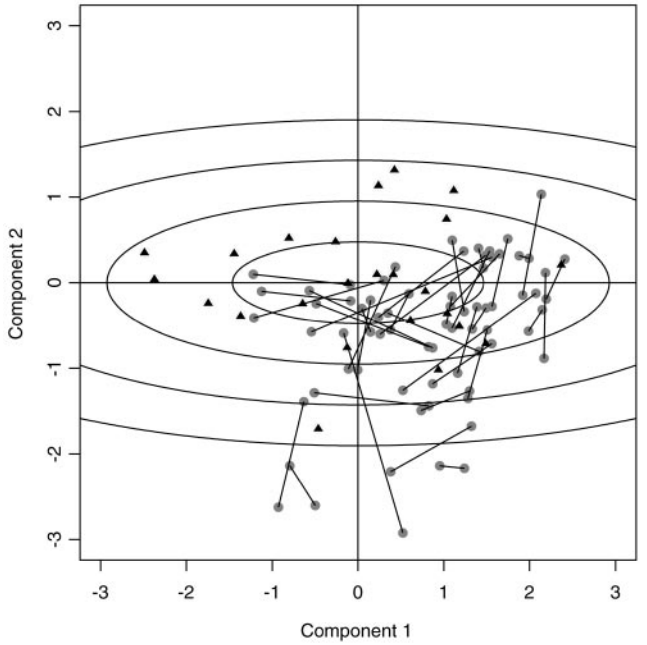

Fig. 4. Projection of new spectra into the space of the first two principal components.

Duplicate spots from new experiments (*gray dots connected by lines*) are usually about the same distance from the center of the original experiments (*black triangles)*. Distance from the center is measured using the Mahalanobis distance, which adjusts for the variance; the *ellipses* represent units of Mahalanobis distance.

Discussion

Modern biology and medicine are changing rapidly in response to the introduction of new high-throughput technologies. Gene expression microarrays (oligonucleotide and cDNA), tissue microarrays, array-based comparative genomic hybridization, and proteomics are altering the research landscape at a tremendous pace. These technologies are being used to screen thousands of potential markers of disease status, progression, or response to therapy. Pressure is growing to find ways to use these technologies as practical tools in the clinics, not just as research instruments. If we are to achieve this goal, it is vital that we develop stringent QC measures to ensure the reliability of the data.

Petricoin et al. *(1)* have described their QC procedure, which relies on multiple QC spots on each ProteinChip array. However, their method relies on detecting the existence of a specified set of only 10 or 15 peaks. Qu et al. *(10)* also spotted two QC spots on each array and used seven peaks in the QC spectra to assess reproducibility and estimate the CV. When trying to apply similar ideas to our spectra, we found that relying on binary indicators (existence or nonexistence) for a small number of peaks was not stable; it seemed to be sensitive to several arbitrary parameters, including the signal-to-noise threshold used to identify peaks, the number of original QC samples in which the peak was required to occur, and the number of QC peaks that were required to be detected. In light of our finding that the heights of peaks are correlated, we also found it difficult to isolate a small number

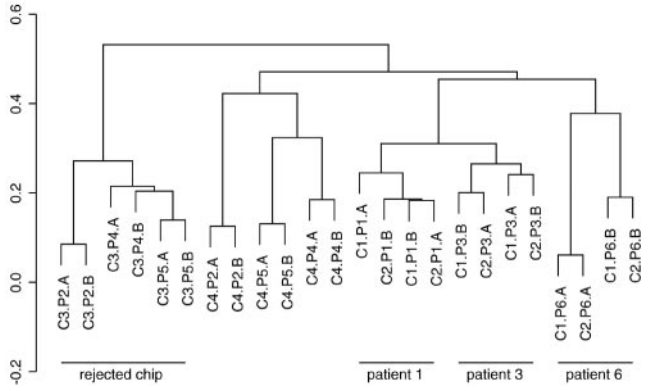

Fig. 5. Hierarchical clustering dendrogram.

Six patients (*P1–P6*) each provided two independent biological samples (*A* and *B*). Each sample was analyzed on two separate ProteinChip arrays (*C1–C4*), producing a total of 24 proteomic profiles. Clustering was based on correlations between the standardized intensities of 356 aligned peaks in the normalized, baseline-corrected spectra. Replicate samples from three patients (P1, P3, and P6) on arrays that passed QC are clustered according to the known biology. Samples on one chip (C3) that failed QC are clustered on a separate branch of the dendrogram instead of with their replicates.

of reliable peaks that provided independent information about the quality of the experiment. This lack of independence makes it more difficult to develop valid statistical measures of the quality.

By contrast, we believe that our method, which uses the measured spectral intensities at a few hundred points across the entire spectrum, provides a more robust way to assess the quality. Furthermore, the use of PCAs provides a natural way to convert this robust set of measurements into a meaningful independent set of measurements, with an easily derived and plausible statistical measure of the deviation of a spectrum from its expected range of values. A similar application of PCA to the QC of microarray experiments has recently been described *(11)*.

Our QC method uses an initial set of experiments to develop a "base profile", applies PCA to estimate the intrinsic variability, and then projects QC samples from future experiments onto this fixed principal component space. Using this design, we can apply our QC procedure immediately to identify experiments that need to be repeated. However, the method assumes that the first experiments are "better" than later experiments. In a practical sense, this assumption is likely to be correct. The first experiments involve a single sample preparation. Later experiments involve multiple samples, which adds complexity and increases the likelihood that the technician will make a mistake. Thus, the experimental process is more likely to be "in control" during the simpler set of experiments that are performed first.

The method can be modified in several ways. For example, keeping the same experimental design, one can perform PCA using all spectra obtained from the QC sample. In our case, the 95% confidence limits would be computed based on all 96 QC spectra instead of just the first 24. Someone who believed that the first experiments are no better than later experiments would prefer this approach. However, this approach cannot be used immediately to decide whether an experiment must be repeated. It also requires a robust version of PCA to avoid distortions caused by including low-quality spectra *(11)*. Another alternative would be to update the base profile as soon as a new experiment is successfully completed. The new PCA profile is used to assess the quality of the next experiment. This alternative, however, can allow the process to drift gradually out of control. One can monitor process drift by use of multivariate control charts, but these again require the establishment of a stable base profile from a time when the experimental process is known to be in control *(11)*.

Yet another approach would forego the idea of a separate QC sample entirely. All spectra from all experimental samples would be used for PCA. Outliers would be detected by setting appropriate confidence limits on the Mahalanobis distance in principal component space. This method could potentially detect individual spot outliers, not just entire chips. It would require robust PCA for the same reason discussed above. It would also require careful attention to randomizing the samples to ensure that all types of experimental samples are equally likely to be found on any given ProteinChip array.

The ideal QC method for proteomic data would ensure the quality of individual spectra, not just the quality of entire chips. To achieve this goal, we would need "gold standard" references for the ProteinChip array surfaces and a standard biological reference material that could be used across laboratories and institutions. Absent those, our method is only one part of a comprehensive approach to QC that should be rigorously applied throughout the entire experimental and analytical process. The quality of protein samples should be checked independently before they are applied to an array. Individual spots and arrays should be tested before they are used. Modifications such as those proposed in the previous paragraph would assess the quality of individual spectra after they are collected. We believe that our method makes substantial contributions toward the goal of developing this comprehensive approach to QC.

We found that use of a different number of principal components for QC can change the results of the analysis. In our experiments, we used the first six principal components. The statistical literature describes several methods for deciding how many principal components to retain *(12)*. The simplest method chooses a target percentage (typically 80% or 90%) for the amount of variance explained. Slightly more sophisticated methods urge retaining components that explain more than the average amount of variance. For our data set, both methods select the first six components, which explain 80% of the variance. In point of fact, we examined the results, using every possible number of principal components. We found that the number of individual spectra rejected (by exceeding the 95% confidence limit) grew with the number of components and leveled off at about six principal components. Moreover, except for minor fluctuations, presumably attributable to spectra close to the threshold, the set of rejected spectra did not change significantly when we used any number of principal components between 6 and 24. Nevertheless, we feel that choosing the correct number of principal components remains an important question for future applications of this method.

Our ANOVA, performed on the initial experiments when the process was most likely to be in control, showed that spot-to-spot variation was larger than day-to-day variation, which was much larger than chip-to-chip variation. For our experiments with patient samples, however, all eight spots of a chip were used on the same day. Thus, the goal of our QC procedure is to identify problems with the experiments performed on one chip on 1 day. Because the spots are variable even under controlled circumstances, however, we were unwilling to throw away the data from an entire chip on the basis of a single bad spot. Thus, we included two QC spots on each chip, and we rejected only the data from a chip if both Mahal-

anobis distances were large enough to exceed a significance threshold.

Our QC procedure was liberal about including peaks, only requiring them to be detected in 2 of the 24 original QC spectra. The ability to find peaks is sensitive to the signal-to-noise threshold used for peak detection. Visual inspection of the identified peaks strongly suggested that all peaks detected at least 10 times, and many peaks detected at least 5 times, represented true peaks in almost all spectra. Of course, that means that some of the "peaks" retained for the QC procedure were caused by random fluctuations that barely exceeded the noise. By including such putative, rarely seen peaks, we allow for the possibility that a newly acquired spectrum might differ from the old spectra by including peaks at several locations where we expect (on average) to see nothing. This could be just as bad as a spectrum that is missing the peaks we are certain should be there. For this reason, we decided that including peaks that had only been detected twice was valuable.

The processing of raw proteomics spectra involves several interrelated problems: noise estimation, baseline correction, and peak finding. Most existing software products proceed through the steps in this order; this is true, in particular, of the ProteinChip software that accompanies the PBSII that was used for the proteomics experiments described here *(8)*. This approach suffers from some major drawbacks: (*a*) it can be difficult to distinguish noise from peaks of low intensity, especially when they are confounded with rapidly changing baseline values; and (*b*) it can be difficult to identify the baseline accurately in spectra produced from serum samples, where we expect to find a large number of peaks. For these reasons, we have developed an alternative method that interweaves the peak finding and baseline correction steps, putting our initial emphasis on the peak-finding step. Simply, we believe that better results can be obtained by making an initial tentative identification of peaks that can be removed before estimating and subtracting the baseline.

In this report, we have demonstrated the potential utility of a new, robust method for monitoring the quality of SELDI proteomics data from NAF samples. We anticipate that this method could be useful for SELDI experiments using other kinds of biological samples. It should also work for other time-of-flight proteomics experiments, provided only that the surface used to conduct the experiment permits the researcher to process multiple samples, including, in particular, the necessary QC samples.

The breast is a unique organ in that its microenvironment can be readily accessed and evaluated by aspiration of fluid from the nipple. Ductal fluids contain large amounts of protein. Recent advances in proteomics, automated mass spectrometry, and the bioinformatics described in this study have provided the necessary tools to utilize ductal fluids from breast cancer patients and other complex biological fluids for biomarker discovery. Toward this end, we are currently using these techniques for analyses of paired ductal fluid samples from patients with unilateral breast cancer in our ongoing prospective investigation designed to identify unique molecules associated with carcinogenesis and progression of this disease.

We are grateful to Sylvie Marcy for assistance with specimen and data collection. This investigation was supported in part by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, by a generous donation from the Michael and Betty Kadoorie Foundation to the Cancer Genomics Core Program, and by the Institutional Research Grant Program for Clinical, Translational, and Population-Based Projects funded through The University of Texas M. D. Anderson Cancer Center.

References

1. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
2. Dooley WC, Ljung BM, Veronesi U, Cazzaniga M, Elledge RM, O'Shaughnessy JA, et al. Ductal lavage for detection of cellular atypia in women at high risk for breast cancer. J Natl Cancer Inst 2001;93:1624–32.
3. Wrensch MR, Petrakis NL, Miike R, King EB, Chew K, Neuhaus J, et al. Breast cancer risk in women with abnormal cytology in nipple aspirates of breast fluid. J Natl Cancer Inst 2001;93:1791–8.
4. Zhao Y, Verselis SJ, Klar N, Sadowsky NL, Kaelin CM, Smith B, et al. Nipple fluid carcinoembryonic antigen and prostate-specific antigen in cancer-bearing and tumor-free breasts. J Clin Oncol 2001;19:1462–7.
5. Evron E, Dooley WC, Umbricht CB, Rosenthal D, Sacchi N, Gabrielson E, et al. Detection of breast cancer cells in ductal lavage fluid by methylation-specific PCR. Lancet 2001;357:1335–6.
6. Kuerer HM, Goldknopf IL, Fritsche H, Krishnamurthy S, Sheeta EA, Hunt KK. Identification of distinct protein expression patterns in bilateral matched-pair breast ductal fluid specimens from women with unilateral invasive breast cancer: high-throughput biomarker discovery. Cancer 2002;95:2276–82.
7. Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. Electrophoresis 2000;21:1164–77.
8. Fung ET, Enderwick C. ProteinChip clinical proteomics: computational challenges and solutions. Biotechniques 2002;32 Suppl: 34–8,40–1.
9. Sartorius OW, Smith HS, Morris P, Benedict D, Friesen L. Cytologic evaluation of breast fluid in the detection of breast disease. J Natl Cancer Inst 1977;59:1073–80.
10. Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin Chem 2002;48: 1835–43.
11. Model F, Konig T, Piepenbrock C, Adorjan P. Statistical process control for large scale microarray experiments. Bioinformatics 2002;18(Suppl 1):S155–63.
12. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. Reading MA: Academic Press, 1979:213–54.