Clinical Chemistry 51:1
102–112 (2005)

Evaluation of Serum Protein Profiling by Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry for the Detection of Prostate Cancer: I. Assessment of Platform Reproducibility

O. John Semmes,[1*] Ziding Feng,[2] Bao-Ling Adam,[1] Lionel L. Banez,[3] William L. Bigbee,[4] David Campos,[5] Lisa H. Cazares,[1] Daniel W. Chan,[6] William E. Grizzle,[7] Elzbieta Izbicka,[5] Jacob Kagan,[8] Gunjan Malik,[1] Dale McLerran,[2] Judd W. Moul,[3] Alan Partin,[6] Premkala Prasanna,[3] Jason Rosenzweig,[6] Lori J. Sokoll,[6] Shiv Srivastava,[3] Sudhir Srivastava,[8] Ian Thompson,[9] Manda J. Welsh,[4] Nicole White,[6] Marcy Winget,[2] Yutaka Yasui,[2] Zhen Zhang,[6] and Liu Zhu[7]

Background: Protein expression profiling for differences indicative of early cancer has promise for improving diagnostics. This report describes the first stage of a National Cancer Institute/Early Detection Research Network-sponsored multiinstitutional evaluation and validation of this approach for detection of prostate cancer.

Methods: Two sequential experimental phases were conducted to establish interlaboratory calibration and standardization of the surface-enhanced laser desorption (SELDI) instrumental and assay platform output. We first established whether the output from multiple calibrated Protein Biosystem II SELDI-ionization time-of-flight mass spectrometry (TOF-MS) instruments demonstrated acceptable interlaboratory reproducibility. This was determined by measuring mass accuracy, resolution, signal-to-noise ratio, and normalized intensity of three *m/z* "peaks" present in a standard pooled serum sample. We next evaluated the ability of the calibrated and standardized instrumentation to accurately differentiate between selected cases of prostate cancer and control by use of an algorithm developed from data derived from a single site 2 years earlier.

Results: When the described standard operating procedures were established at all laboratory sites, the across-laboratory measurements revealed a CV for mass accuracy of 0.1%, signal-to-noise ratio of ~40%, and normalized intensity of 15–36% for the three pooled serum peaks. This was comparable to the intralaboratory measurements of the same peaks. The instrument systems were then challenged with sera from a selected group of 14 cases and 14 controls. The classification agreement between each site and the established decision algorithm were examined by use of both raw peak intensity boosting and ranked peak intensity boosting. All six sites achieved perfect blinded classification for all samples when boosted alignment of raw intensities was used. Four of six sites achieved perfect blinded classification with ranked intensities, with one site passing the criteria of 26 of 28 correct and one site failing with 19 of 28 correct.

[1] Department of Microbiology & Molecular Cell Biology, Virginia Prostate Center, Eastern Virginia Medical School, Norfolk, VA.

[2] Fred Hutchison Cancer Center, Seattle, WA.

[3] Center for Prostate Disease Research, Department of Surgery, Uniformed Services University of the Health Sciences, Rockville, MD.

[4] University of Pittsburgh Cancer Institute, Hillman Cancer Center, Pittsburgh, PA.

[5] Cancer Therapy and Research Center, Institute for Drug Development, San Antonio, TX.

[6] Department of Pathology, Johns Hopkins Medical Institutes, Baltimore, MD.

[7] Department of Pathology, University of Alabama at Birmingham, Birmingham, AL.

[8] Cancer Biomarkers Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD.

[9] Department of Medicine, University of Texas Health Sciences Center, San Antonio, TX.

*Address correspondence to this author at: Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, 700 W. Olney Rd., Norfolk, VA 23507. Fax 757-446-5766; e-mail semmesoj@evms.edu.

Received June 22, 2004; accepted September 21, 2004.

Previously published online at DOI: 10.1373/clinchem.2004.038950

Conclusions: **These results demonstrate that "between-laboratory" reproducibility of SELDI-TOF-MS serum profiling approaches that of "within-laboratory" reproducibility as determined by measuring discrete *m/z* peaks over time and across laboratories.**

Despite three decades of intense effort in developing new therapies to improve the survival of cancer patients, the results are modest with the exceptions being selected childhood cancers *(1–4)*. The majority of patients diagnosed as having cancer are late stage. For example, 72% of lung cancer patients, 57% of colorectal cancer patients, and 34% of breast cancer patients in the United States are diagnosed at late stage with regional or distant dissemination of the cancer cells *(2)*. On the other hand, when these cancers are diagnosed at early stage and are organ confined, the survival rate exceeds 85% *(1, 2)*. Clearly, with the current available therapies and treatments, only improvements in early detection of cancer will lead to improvements in cancer survival. It is therefore crucial to develop high-throughput noninvasive or minimally invasive tests to diagnose cancer at early stages.

Several laboratories have demonstrated the feasibility of using mass spectrometric proteomic pattern analysis for the diagnosis of several categories of tumors, including ovarian, breast, lung, pancreas, and prostate cancer *(5–24)*. In one of the earliest contributions, investigators from Eastern Virginia Medical School (EVMS)[10] and the Fred Hutchison Cancer Research Center (FHCRC) demonstrated that surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) protein profiles may have clinical utility as a diagnostic tool for the early detection of prostate cancer, including prostatic adenocarcinoma (PCa) *(6, 8)*. Specifically, using just nine differentially expressed protein peaks, the assay achieved an overall PCa case/control correct classification rate >90%. Also of interest was the high specificity, 97%, which is significantly better than that of prostate-specific antigen (PSA) and would therefore be of obvious clinical use for PCa screening. This is particularly important given the recent evidence that current PSA screening practices do not detect the majority of prostate cancers, including those that are high grade *(25)*.

Recently, there has been considerable controversy concerning the SELDI profiling approach *(26)*; the most publicized of which have been several commentaries by Diamandis and the analysis of publicly available data posted by Petricoin and coworkers. The major concern raised by Diamandis *(27, 28)* has been whether the SELDI-based approaches are reproducible, whereas the concerns raised by Sorace and Zahn *(29)* and Baggerly et al. *(30)* deal with study design bias. There is also a great deal of confusion and controversy as to whether the use of high-throughput proteomic techniques such as SELDI can improve the early detection of prostate cancer, including PCa *(6, 9, 15, 28, 31)*. In addition, clinical issues as to whether SELDI detects equally well both high-grade (Gleason score ≥7) and low-grade (Gleason score ≤6) cancers of the prostate remain to be determined. Similarly, a very important comparison will be the use of SELDI-TOF-MS in the early detection of PCa compared with the present implementation and future use of assays for PSA. All of these issues can and should be addressed in any appropriately designed validation study *(32)*.

It is the goal of this collaborative project—the Early Detection Research Network (EDRN)-Prostate-SELDI Investigational Collaboration (EPSIC)—to use state-of-the-art protein profiling technology to develop and validate high-throughput screening methods for the early detection of PCa. A successful validation study of SELDI profiling is a necessary first step toward demonstration of clinical viability of the assay. In addition, we have included an initial first phase to determine reproducibility of the SELDI assay as a specific response to the question of platform reproducibility. This report describes the methods, results, and pitfalls encountered in the initial assay reproducibility assessment phase of our analysis.

Materials and Methods

STUDY DESIGN

Phase I of the EPSIC validation study consisted of two major parts reported here, A and B, which were each to be completed sequentially and successfully before moving to subsequent parts. The primary purpose of the initial phase (phase 1A) of the validation was to determine whether SELDI-TOF-MS instruments could be accurately calibrated and the output standardized across all participating institutions (see Table 1). The primary purpose of the second stage (phase 1B) was to determine whether each calibrated and standardized instrument could consistently differentiate between the sera of previously characterized PCa and non-PCa individuals. Specifically, in the first stage we calibrated the instruments by use of an established set of protocols and then standardized the instrument output with respect to three prominent *m/z*"peaks" present in a pooled serum sample [quality control (QC)]. This process was iterative and repeated until specifications were met. The second phase was a single blinded test of each calibrated and standardized instrument with respect to the ability to reproducibly measure previously selected "diagnostic" *m/z* peaks. The chosen diagnostic peaks were identified from analysis of a large 1000+ patient cohort, which we will refer to as EVMS-2002. The criteria for patient inclusion in this study gave three major groups: PCa, benign disease, and no

[10] Nonstandard abbreviations: EVMS, Eastern Virginia Medical School; FHCRC, Fred Hutchison Cancer Research Center; PCa, prostatic adenocarcinoma; SELDI-TOF-MS, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry; PSA, prostate-specific antigen; EDRN, Early Detection Research Network; EPSIC, Early Detection Research Network-Prostate-SELDI Investigational Collaboration; QC, quality control; and S/N, signal-to-noise ratio.

Table 1. Acceptance criteria.[a]

Protein	S/N	Resolution, *m/z*	Intensity[b]
Insulin	NA[c]	600	
IgG	700	NA	
QC			
Peak 1 (*m/z* 5910 ± 0.2%)	>40	>400	24 (9)
Peak 2 (*m/z* 7773 ± 0.2%)	>80	>400	37 (5)
Peak 3 (*m/z* 9297 ± 0.2%)	>80	>400	30 (3)

[a] Quality control assessment of the Protein Biosystem II is based on mass designation, S/N, resolution, and normalized intensity. The table provides the values required for the QC protein peaks used in phase 1A for a site to proceed to phase 1B.

[b] Mean (SD) normalized intensity.

[c] NA, not applicable.

evidence of disease with stratified PSA concentrations. The results from this study, including the peak values, are not revealed so that our second phase study remains blinded. The EVMS-2002 study is separate from the previously published EVMS study *(6)* and was designed specifically for to meet the overall EPSIC objectives. We then used both raw intensity and ranked intensity approaches to subselect the diagnostic peaks. The final stage of phase I, which is in progress, will assess the concordance of each site when challenged with new PCa cases and controls from multiple clinical sites. A full description of the complete EPSIC study design has been described by us elsewhere *(32)*.

interlaboratory calibration and standardization of instrument output

SELDI-TOF-MS is a modification of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. The principal difference is the addition of an affinity chromatographic step incorporating specific chemical modification of the desorption surface. Thus, sample complexity is reduced "on-chip" in a convenient, easily manipulatable format before laser desorption. The specific system used for these studies is the Protein Biosystem II model outfitted with the metal affinity (IMAC3-Cu) ProteinChip®. Each of the six laboratory sites was equipped with a Protein Biosystem II (Ciphergen Biosystems Inc.) and a Biomek® 2000 Workstation liquid-handling robot (Beckman Instruments), which included an eight-channel pipette and wash tool and a custom-configured mixer/shaker. The robot is adjusted to prepare the samples and spot them onto the Ciphergen ProteinChip arrays by use of the Ciphergen 12-array Bioprocessor® unit. Sinapinic acid matrix addition was also performed robotically. These detailed preparative protocols have been described elsewhere *(6, 19)*, and the complete standard operating procedure is included in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol51/issue1/.

The synchronization and standardization process for phase IA of the study was performed as follows: Aliquots of pooled normal human sera prepared at EVMS were used as QC samples. This QC sample consists of pooled serum obtained from 360 healthy individuals (197 women and 163 men), giving almost 2 L of serum. Serum from each individual was collected by venipuncture into a 10-mL SST Vacutainer™ Tube. Blood was allowed to clot at room temperature for 30 min, and the tubes were centrifuged at 1500*g* for 10 min. Each individual serum sample was then decanted and pooled into a 3-L beaker on ice. The pooled serum was separated into 0.4-mL aliquots and stored at −80 °C. Each aliquot of the QC pool has therefore undergone only one freeze-thaw cycle. Spectra obtained for these QC samples at EVMS were used as benchmarks for the participating sites, and a QC sample is run on one randomly chosen spot of every chip processed. Protocols and a checklist of required criteria were sent to each site, and technical staff from EVMS and Ciphergen Biosystems coordinated a visit to each institution to standardize the Protein Biosystem II instruments and prepare the robotics for serum processing. The QC serum samples were assayed, and the resulting data were used to evaluate the performance of both the Biomek 2000 robotic sample-processing station and the Protein Biosystem II SELDI-TOF-MS instrument at each site. Instrument calibration of the spectra was performed externally with the All-in-1 peptide calibrator (Ciphergen), which contains seven peptides in the mass range of 1084.25–7033.61 Da. Instrument calibration was performed weekly, and the most recent calibration equation was applied.

After routine instrument performance was tested, standardization of output was accomplished by adjusting the laser intensity, detector voltage, and detector sensitivity of each instrument so that three consistently present protein peaks in the QC serum (*m/z* 5910, 7773, and 9297 ± 0.2%) were displayed to specific criteria (see Table 1). Additionally, the resolution values for all three peaks were required to be >400, with signal-to-noise ratios (S/N) ≥40 for *m/z* 5910 and ≥80 for *m/z* 7773 and 9297. When these criteria were met, each onsite researcher then performed an evaluation phase. This consisted of running two additional Bioprocessors (12 ProteinChips/each) of QC and sending the data to EVMS for examination and initial analysis. The same raw data were also sent to the FHCRC for final analysis.

During the course of the initial validation phase, the QC pool was assayed weekly. If any set of QC spectra did not meet the criteria specified, instrumental data collection settings were adjusted. If the QC spectra still did not meet the expected criteria, robot protocols were checked and adjusted. If this did not resolve the problem, the Protein Biosystem II instrumentation was then checked by the Ciphergen engineer.

samples and data collection

After phase 1A standardization of the SELDI-TOF-MS system at a site, each site received 14 PCa and 14 non-PCa sera from EVMS. All serum samples used in this study were collected and assayed under a protocol approved by the respective Institutional Review Boards for each laboratory site. Criteria used to select the 14 PCa and 14 non-PCa samples from 186 PCa patients and 219 age-matched healthy men included that (*a*) the spectra show no signs of protein degradation, (*b*) at least 1 mL of serum remain in the serum bank, and (*c*) samples were correctly classified by the current classifier. We did not use samples that were marginally classified because the objective of phase 1B was platform validation and not biological validation of the previously derived classifier. The operator laboratory sites were blinded to the mass locations of the three diagnostic peaks. Samples were coded and analyzed blindly by each of the six laboratories, including EVMS. Each of the six sites prepared these samples on IMAC3-Cu ProteinChips, using their Biomek robotic system, and analyzed them on the SELDI-TOF-MS system. No effort was made to ensure that the ProteinChips used were from the same lot, and multiple lots were used within and across laboratory sites. All sites sent their primary data and processed spectra to the coinvestigators at the FHCRC, the Data Management and Coordinating Center of the EDRN, which analyzed each site's ability to identify the peaks required for accurate analysis and classification of these samples.

DATA ANALYSIS

The classifier used in this study was constructed by use of boosting (boosting logistic regression and boosting decision tree). The details have been described previously *(8, 33)*. The peak identification and alignment method used has also been described *(34)*. Three diagnostic peaks were fixed for this classifier and used in phase 1B. The maximum intensity value in each of the three windows centered at the diagnostic peaks, with a half window width of ±0.2% of mass value, was used for each triplicate, and the median intensity value was selected. The median intensity value for each diagnostic peak was then used in the classifier to calculate a predication score. If the prediction score value was >0, the predictive probability for this sample being a cancer was greater than that for being a control, and the sample would be classified as a cancer; otherwise, it would be classified as a control. Scores of all 28 samples were calculated for each site by use of this fixed classifier, and the results were compared with the scores calculated on the original spectra generated 17 months earlier at the EVMS site.

Results

PHASE IA: CALIBRATION OF SELDI-TOF-MS INSTRUMENTATION AND STANDARDIZATION OF SPECTRAL OUTPUT FROM MULTIPLE SITES

The SELDI spectral output from each validation site was standardized by examining and adjusting the spectral output from the QC serum samples. Two Bioprocessors or 192 spots were analyzed for each site. Representative QC spectra from each validation site (Fig. 1) show the three most prominent peaks present in the spectra used for instrumental output standardization. All spectra were compiled for each site that passed the criteria listed in Table 1. Data from each site were collected after weekly internal calibration and had been normalized. The variations in peak *m/z*, peak intensity, peak resolution, and S/N of the three peaks in the QC spectra were recorded. The individual results from each laboratory according to the specified QC peak criteria are shown in Table 2, and Table 3 displays the across-laboratory variation. CV values for the mass designation (*m/z*) of each peak ranged from 0.03% to 0.09%. This exceeds the manufacturer's specification for mass accuracy for the Protein Biosystem II instrument, which is 0.1% with instrument calibration, with the note that because the true masses for these three unidentified proteins are unknown, the long-run mean peak locations were used as the "true" mass values. The CVs for S/N were 34–40%. However, this value excluded the S/N for the University of Pittsburgh Cancer Institute

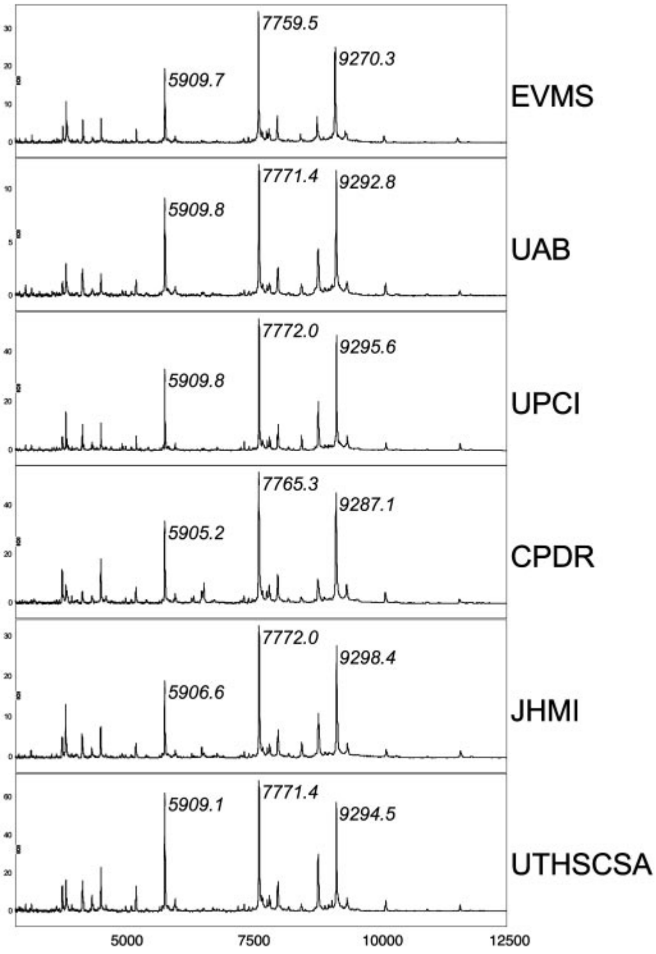

Fig. 1. Representative SELDI spectra.

Shown are SELDI-TOF-MS IMAC3-Cu spectra for the QC serum from each validation laboratory site. The spectra were processed with baseline subtraction and normalization. The three peaks used for instrument synchronization are labeled. *UAB*, University of Alabama at Birmingham; *UPCI*, University of Pittsburgh Cancer Institute; *CPDR*, Walter Reed Army Medical Center, Center for Prostate Disease Research; *JHMI*, Johns Hopkins Medical Institutions; *UTHSCSA*, University of Texas Health Science Center at San Antonio.

Table 2. Intralaboratory variability of QC pool (96 spectra).

	Intensity	Mass, *m/z*	S/N	Resolution, *m/z*		Intensity	Mass, *m/z*	S/N	Resolution, *m/z*
EVMS					CPDR				
Peak 1					Peak 1				
Mean	24.1	5910.9	55.8	519.8	Mean	24.9	5911.4	69.1	383.4
SD	8.5	1.9	23.4	45.8	SD	5.4	1.7	27.5	30.4
CV, %	35	0.03	42	9	CV, %	22	0.03	40	8
Peak 2					Peak 2				
Mean	37.3	7773.6	124.1	525.4	Mean	33.8	7771.9	131.6	450.5
SD	4.9	2.3	45.7	39.7	SD	3.2	2.1	34.4	25.8
CV, %	13	0.03	37	8	CV, %	9	0.03	26	6
Peak 3					Peak 3				
Mean	29.7	9297.8	134.6	407.4	Mean	32.1	9293.2	181.4	423.9
SD	3.4	2.7	42.5	35.0	SD	4.5	2.6	48.6	29.8
CV, %	11	0.03	32	9	CV, %	14	0.03	27	7
UAB					UTHSCSA				
Peak 1					Peak 1				
Mean	26.6	5905.1	60.5	338.4	Mean	36.2	5912.6	73.7	402.1
SD	10.8	5.0	31.9	79.0	SD	5.7	2.8	18.5	42.1
CV, %	41	0.08	53	23	CV, %	16	0.05	25	10
Peak 2					Peak 2				
Mean	34.4	7764.9	108.4	432.7	Mean	31.4	7775.2	93.1	458.7
SD	7.9	5.8	47.1	69.7	SD	4.2	3.7	26.0	55.5
CV, %	23	0.07	43	16	CV, %	13	0.05	28	12
Peak 3					Peak 3				
Mean	31.9	9286.2	136.8	406.9	Mean	31.3	9298.7	136.3	450.6
SD	6.4	7.3	49.3	59.9	SD	3.8	4.3	37.1	53.1
CV, %	20	0.08	36	15	CV, %	12	0.05	27	12
UPCI					JHMI				
Peak 1					Peak 1				
Mean	28.3	5901.7	536.3	589.1	Mean	20.6	5898.6	57.9	452.5
SD	9.3	5.3	215.8	48.0	SD	8.8	7.9	22.6	69.5
CV, %	33	0.09	40	8	CV, %	43	0.13	39	15
Peak 2					Peak 2				
Mean	37.6	7760.6	683.8	601.1	Mean	39.3	7758.6	166.5	515.8
SD	4.7	7.1	146.3	65.5	SD	7.7	9.4	57.2	59.0
CV, %	13	0.09	21	11	CV, %	20	0.12	34	11
Peak 3					Peak 3				
Mean	32.0	9281.8	571.5	538.2	Mean	27.9	9281.4	172.1	381.6
SD	3.5	8.7	112.5	67.9	SD	4.6	11.0	51.7	58.5
CV, %	11	0.09	20	13	CV, %	16	0.12	30	15

[a] CPDR, Walter Reed Army Medical Center, Center for Prostate Disease Research; UAB, University of Alabama at Birmingham; UTHSCSA, University of Texas Health Science Center at San Antonio; UPCI, University of Pittsburgh Cancer Institute; JHMI, Johns Hopkins Medical Institutions.

site because their instrument was the Protein Biosystem IIc model, which has higher resolution and S/N capabilities than the Protein Biosystem II instruments installed at the other EPSIC sites. The across-laboratory CV of resolution for each QC peak was 8–23%. Intensity variation for the smallest QC peak (peak 1, *m/z* 5906) was the highest, whereas the largest peak (peak 3, *m/z* 9289) had the lowest variation. The variations in the intensities of the three peak for all validation sites combined were 36% for peak 1, 17% for peak 2, and 15% for peak 3. These variations in intensity are consistent with the individual site variation for each peak observed in Table 2. Thus, the combined across-site variability was comparable to the observed single-site variability.

The entire phase IA, including data analysis, was completed in ~6 months. Table 1 in the online Data Supplement shows the timelines for site visits, data collection, and analysis. Once the initial observations and evaluation of the QC spectra were completed, each validation laboratory was then directed to continue to the next phase of the study, phase IB.

frequent and accurate instrument calibration is essential for reproducibility

The values for both *m/z* and peak intensity are critical components to the construction of the SELDI-TOF-MS spectral profile. Because consistency for these measures would be critical to reproducible performance, we exam-

Table 3. Interlaboratory variability of QC pool (96 spectra).

	Mass, *m/z*	Intensity	S/N[a]	Resolution
Peak 1				
Mean	5906.5	26.6	61.8	460.7
SD	6.7	9.7	26.5	107.7
CV, %	0.11	36	40	23
Peak 2				
Mean	7768.6	35.9	123.1	505.5
SD	8.4	6.3	49.6	82.8
CV, %	0.10	17	40	16
Peak 3				
Mean	9289.2	31.0	147.7	439.3
SD	9.9	4.7	49.6	77.4
CV, %	0.11	15	34	18

[a] Values calculated without University of Pittsburgh values, which are much higher because Protein Biosystem IIc instrumentation was used.

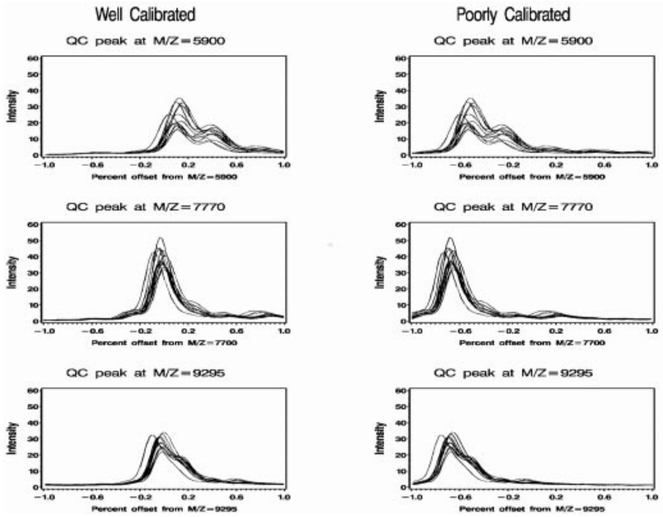

Fig. 2. Impact of proper calibration.

Adequately calibrated (*left panels*) and poorly calibrated (*right panels*) spectra for three internal QC peaks from one site are shown. The poor calibration data were generated by use of a 7-in-1 peptide calibrator in which two peptides were poorly measured. The adequate calibration data were generated with a 7-in-1 peptide calibrator in which all peptides were accurately measured.

ined the impact of accurate mass designation and spectral intensity on cross-site agreements. For this experiment, we compared the spectral data from each site under two calibration scenarios. In the first scenario, the spectral output was correctly calibrated by optimized analysis of a 7-in-1 peptide calibrator; in the second, the spectral output from the 7-in-1 peptide calibrator was poorly resolved. The poor calibration occurred because two of the peptides peaks from the 7-in-1 peptide solution had observed mass values disparate from theoretical, leading to a poorer calibration. The right-hand panels of Fig. 2 show QC peaks from the poor calibration attempt. The left-hand panels show QC peaks based on the 7-in-1 peptide calibrator that demonstrated adequate calibration. When the poor calibration data were used, the mass designation values where dramatically shifted outside the acceptable range. The misclassification rate based on the poor calibration was 93%, whereas use of the adequate calibration led to a misclassification rate of 0%. If the calibration was performed with the 7-in-1 peptide calibrator but the two poorly resolved peptide peak measures were omitted from the calibration, the misclassification error was 0% as well (data not shown).

PHASE IB: CLASSIFICATION OF PCa CASE/CONTROL SAMPLES

The second step in the EPSIC phase I design was to determine whether, after instrument calibration and output standardization, the separate sites could achieve comparable correct classification rates when challenged with the same previously characterized 14 PCa and 14 control samples. The 14 PCa and 14 control sera were provided by EVMS and selected from the larger EVMS-2002 pool originally used to construct the classifier. These 28 samples were chosen based on being easiest to separate by the classifier. The direct assessment of the between-site agreement was to apply the classifier previously developed from EVMS-2002 data gathered 2 years previously to the current data developed at each site. Our pass criterion for this phase IB study was that at least two predesignated sites, EVMS and the University of Alabama at Birmingham (the EDRN-designated Biomarker Discovery Laboratory and the Biomarker Validation Laboratory, respectively) achieve correct classification of at least 26 of 28 samples. Thus, the challenge was to both reproduce the classification results from 2 years earlier when the same samples were analyzed and to demonstrate that the standardized instruments would match this output. The scores and classifications of the 28 samples for all six sites compared with the original spectra are shown in Fig. 3. For good agreement, the values of the scores for all six sites should be near the values for the original scores and also achieve a high degree of correct classification. As shown in Fig. 3, correct classification was achieved for data from all six sites. In particular, the majority of the values of scores for PCa spectra were near their corresponding original scores (Fig. 3, filled circles). Interestingly, the classification of control samples, although 100% correct within the decision window, demonstrated more variation in their scores within each sample. Also of note, there appears to have been larger intersite variation for control samples; most notably with the scores from two sites, EVMS and Walter Reed Army Medical Center, Center for Prostate Disease Research, moving toward zero.

AMONG-SITE AGREEMENT

We next examined the individual spectral output from each laboratory site for the fine scale agreement at each of the peak values used in the diagnostic algorithm. In Fig. 4, median intensity across three replicates per sample was plotted for a random sample of PCa and controls from the

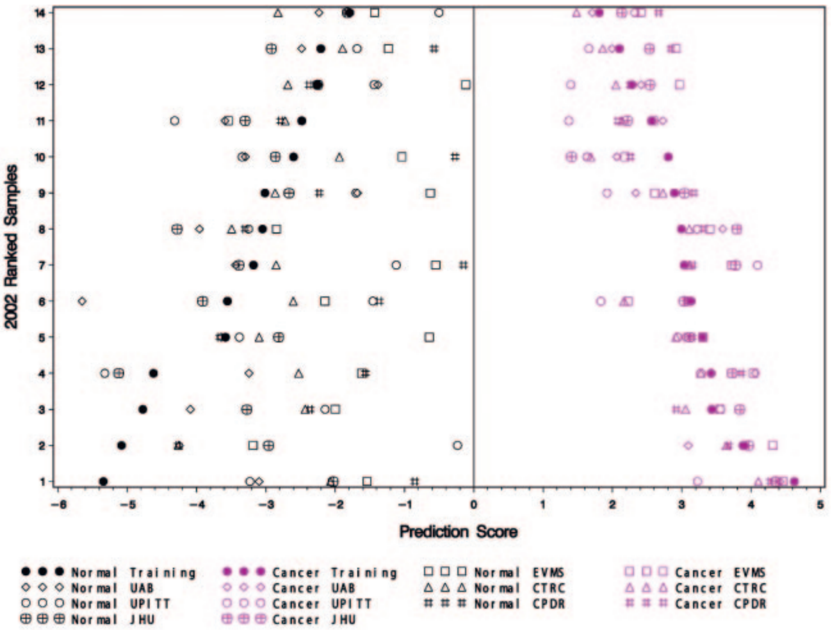

Fig. 3. Interlaboratory prediction scores.

Shown are the plots of the prediction scores of 14 PCa samples and 14 control samples obtained from all six validation laboratory sites and the corresponding scores (*filled circles*) from the original spectra assayed 17 months earlier at EVMS. A score >0 is classified as PCa, and a score <0 is classified as control. *UAB*, University of Alabama at Birmingham; *UPITT*, University of Pittsburgh Cancer Institute; *JHU*, Johns Hopkins Medical Institutions; *CTRC*, University of Texas Health Science Center at San Antonio; *CPDR*, Walter Reed Army Medical Center, Center for Prostate Disease Research.

EVMS-2002 study as representative of the training data set. In addition, Fig. 4 shows the median sample values from the EVMS-2002 study for the 14 PCa and 14 control samples selected for the validation study, together with the diagnostic spectra for the same 28 samples from the six sites. A few points are worth noting. Peak 1, the most dominant diagnostic peak, has the best agreement among the participating laboratories. When we binned the window values, the major features of the peaks, intensities, were captured. Peaks 2 and 3 are weak peaks, and agreement among the laboratories seemed more difficult to achieve for these peaks (see Figs. A and B in the online Data Supplement). Therefore, a classifier based on strong peaks seems more likely to be reproducible across multiple laboratories. All three peaks in all laboratories were shifted to the left when compared with the peaks in the same 28 samples and that in the training samples, indicating that the training sample may have bias in calibration.

We also used a ranked intensity approach to select for diagnostic *m/z* peaks from the larger EVMS-2002 dataset because this approach might be expected to be more robust than the raw intensity method. However, the ranked intensity method led to selection of a major discriminating *m/z* value that corresponded to a "shoulder" peak (see Fig. C in the online Data Supplement). This was in contrast to the separate peaks (dominant *m/z* value within the local spectral neighborhood) used in the raw intensity approach. When the ranked intensity approach was evaluated by use of the data for the 14 cases and 14 controls from each of the six sites, four of the sites achieved perfect classification, one of the sites passed with 26 of 28 correctly classified, and one site failed with 19 of 28 correctly classified (data not shown).

Discussion

The concept of pattern recognition in proteomics is not a recent development, but the current attention being focused on this approach and, specifically, the application of moderately high-throughput methodologies such as SELDI-TOF-MS proteomic profiling, has generated equivalent excitement and concern. Although it would be particularly irresponsible to overlook the potential benefits of this technology in the clinical diagnostic and prognostic arena, we would be equally remiss to accept findings generated from a new technology without rigorous validation. The overall validation plan proposed by the EPSIC is designed to achieve a complete evaluation of protein profiling as a diagnostic tool for PCa. We have structured the EPSIC study to be achieved in carefully designed phases and to incorporate the biomarker validation concepts outlined by Pepe et al. *(35)* and elaborated on in a recent review by Ransohoff *(36)*. Our complete validation plan included examination of the reproducibility of the SELDI-TOF-MS platform as preliminary studies (phase I) before initiation of clinical validation (phases II and III) and has been described in detail elsewhere *(32, 37)*. We would propose that the results from our planned multiinstitutional validation study can be viewed as direct experimental data addressing several of the key questions raised in a recent commentary by Diamandis *(27)*. Specifically, although several meritorious issues were raised by Diamandis, such as the sensitivity to detect low-abundance serum protein markers, a major underlying concern can be summarized as uncertainty with platform reproducibility. In addition, we would also include assay robustness, the demonstration of reproducibility across sites, as a concern of central importance. We feel that completion of rigorously designed and standard-

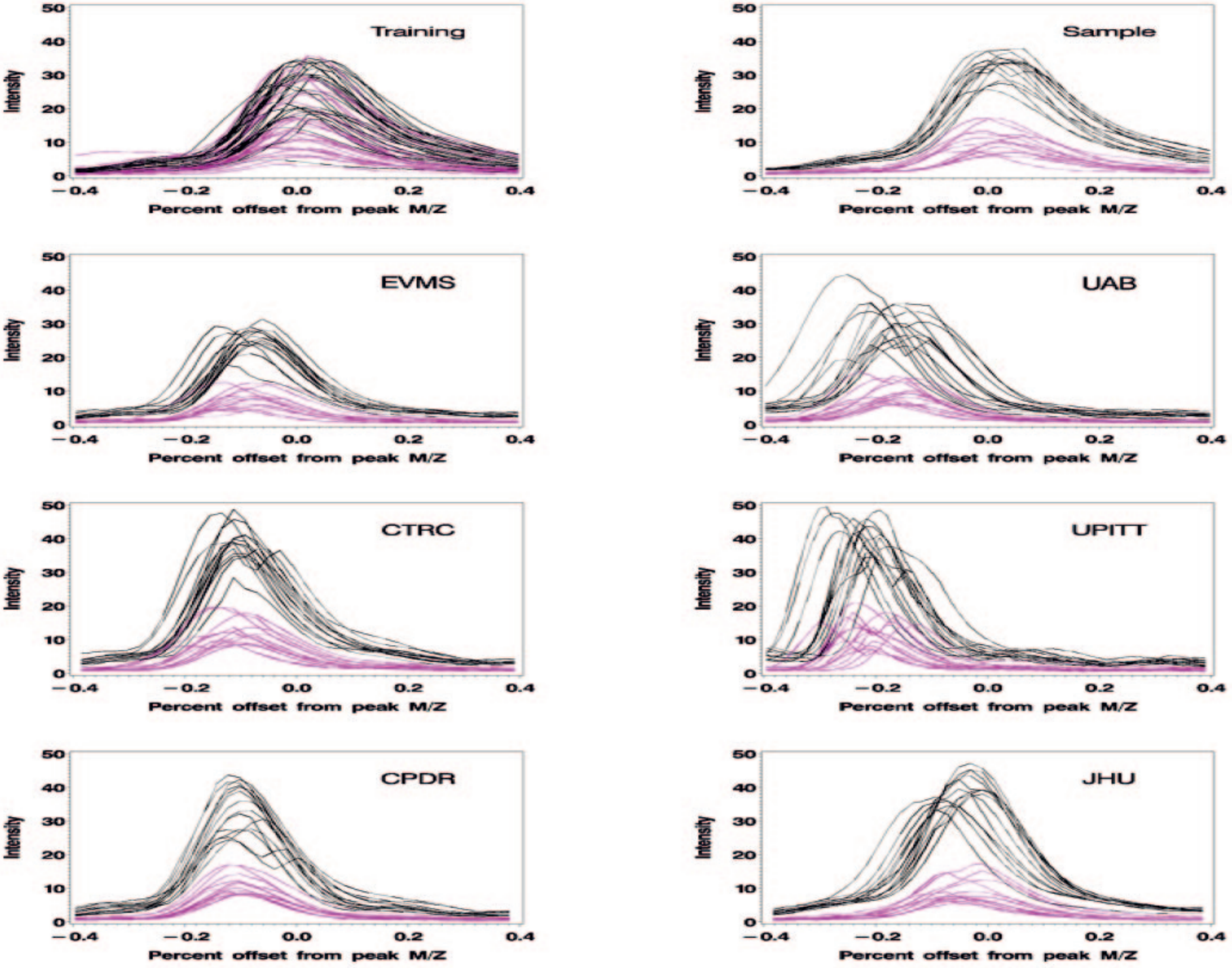

Fig. 4. Interlaboratory peak traces of spectra for diagnostic peak 1.

Sample (*top right*) represents the spectra of the 28 samples (14 cases and 14 controls) measured at EVMS site for this study. *Training* (*top left*) represents a sample of spectra for 28 PCa cases and controls from all training samples run 17 months previously in the EVMS-2002 study. The other *panels* represent spectra obtained from each validation site for the sample set, using the current protocol of calibration and instrument standardization. *Red lines* indicate the peak traces of cases, and *black lines* indicate controls. *CTRC*, University of Texas Health Science Center at San Antonio; *CPDR*, Walter Reed Army Medical Center, Center for Prostate Disease Research; *UAB*, University of Alabama at Birmingham; *UPITT*, University of Pittsburgh Cancer Institute; *JHU*, Johns Hopkins Medical Institutions.

ized studies is the best approach to addressing these concerns, and indeed, none of the concerns being raised are unique to SELDI-TOF-MS profiling and should be important elements of all prevalidation and validation efforts. The preoccupation with protein identification, albeit an important parallel goal, has caused many to lose focus on the important issues of proper study design and validation of the observed proteomic profiles, which we hope to address in the EPSIC.

The first objective evaluation of the SELDI-TOF-MS methodology was for the ability to standardize the output of six Ciphergen Protein Biosystem II instruments. Specifically, we are asking whether the platform and assay conditions can be transferred from an originating laboratory site to other secondary sites. This is an important test because it also requires reproducibility of established constant conditions over time. Thus, regardless of the actual identities of the proteins that give rise to the observed SELDI-TOF-MS peaks, the expression patterns will need to be shown to be reproducibly observed from site to site. To successfully accomplish this first test (phase IA), each site had to adopt strict standard operating procedures. We found it absolutely necessary for each site to monitor continuously, whenever data were gathered, the spectral output of each instrument and adjust it as necessary to achieve the exact display of several peaks consistently found in the QC serum. Indeed, it is obvious from our results that automated software/hardware improvement is needed in this regard to ensure reproducible optimization of instrumental operation. It is also clear that automation of all sample preparation steps is essential to achieving accurate reproducibility. Several instrument parts were also replaced to keep the instruments in optimal performance, and this list has also been included in the online Data Supplement.

Once the instruments at each of the six laboratory sites had been calibrated and the output standardized, they were ready to be challenged by PCa case and control samples. In this test, each instrument was required to assay case/control samples from the same source. These case/control test samples originated from the larger EVMS-2002 sample set that was used to generate the decision algorithm being challenged. Thus, a static decision algorithm and standardized assay approach was being evaluated. Two models were used to initially select the 14 cancers and 14 controls that were to be separated in phase 1B. One model used boosting applied to aligned peak raw intensities, and these results are presented in full. The other model, which we include for comparison, ranked the peak intensities within each sample from 1 to 100. (Note that there are more than 100 peaks to rank; therefore, peak intensities are aggregated into bins by use of the rank approach.) Boosting was then applied to the rank values. Cancer and control cases that were furthest apart according to both models were selected for phase 1B. In phase 1B, we evaluated the performance of both of these classifiers. The classifier built by boosting of the peak intensities had perfect classification for all six sites. The classifier built by boosting of the ranked intensities had perfect classification for four sites, passable classification at one site, and failed our criteria for classification at one site. It is interesting that the inherently more robust ranked intensity approach had more difficulty achieving perfect classification compared with the raw intensity value approach. We believe that this is attributable to the use of a shoulder peak *m/z* value in the final decision structure. This would suggest that complex peak shapes will be more difficult to accurately reproduce and may fail as robust marker events. This result underscores the necessity of defining peak vs "noise" events from the *m/z* spectral data as a required preliminary step.

Although the EPSIC passed the criteria set for phase 1B using either value selection approach, there are several other issues worth discussion. The first is the role of calibration. The process for calibrating the SELDI-TOF-MS instrument is explained in the manufacturer's manual, but the importance of the process is not made very clear. The described approach involves the use of a peptide calibrator, which is run on the SELDI instrument, and the spectral output is used to calibrate the conversion of time of flight to mass/charge (*m/z*) values. We recommend that the peptide calibration equation is always first checked by its residual plot for goodness of fit and then compared with the spectral peaks of the pooled serum for mass designation. An inadequate calibration equation can lead to a significant shift of the *m/z* values for the peak maximum. If this had occurred during our analysis, the sites using this poor calibration would have failed to classify correctly. These poorly calibrated spectra are detected only by careful examination of the profiles of the QC spectral peak locations on the magnitudes of the observed *m/z* shifts of peaks. A cursory visual inspection of the overall spectra will likely miss this important problem and consequently lead to failure of the classification comparisons. We are currently developing analytical software approaches to incorporate external/internal reference peaks to prevent this from occurring.

The optimization/standardization of peak intensities is also important for successful classification. For all three diagnostic peaks, cases have lower intensities than controls, and the diagnostic peak in Fig. 4 is the most dominant diagnostic peak feature for this classifier. Compared with the other sites, spectra obtained at the EVMS site showed lower intensities for the control samples. This may explain the observation that the control scores from the EVMS laboratory site were more shifted toward zero than were the scores from the other sites. In addition, peak 1, the most dominant diagnostic peak, had the best agreement among the sites. When we binned the windows, the major features of the peaks, such as intensities, were captured because maximum values within the bin usually took the values near the peaks. Peaks 2 and 3 are relatively weak peaks, and agreement among the sites was more difficult to achieve for these peaks features (see the online Data Supplement). Therefore, a classifier based on strong peaks seems to have higher likelihood of being reproducible across multiple laboratories. All three peaks in the spectra from all of the sites were shifted to the left when compared with the same peaks in the original training samples, indicating that the training set may possess some bias in calibration.

The Protein Biosystem II instruments at several sites required repair or parts replacement over the course of phases A–C (see Table 2 in the online Data Supplement). In addition, several sites required that the detector voltage be incrementally increased over time to produce QC spectra that consistently passed the specified criteria. This is a result of detector decay over time, which is known to occur in this instrumentation. We recommend that, in addition to continual monitoring of the QC spectra, laboratories should keep log books of usage to anticipate replacement and detector voltage drift events.

In summary, it may be helpful to reconsider the comments of Diamandis *(27, 28)*. Diamandis raised concern that the "discriminating peaks are not consistent either within a group or among groups of individuals". This comment, which was derived from a "metaanalysis" of the published literature, has now been experimentally addressed in our study in which we show clearly that the same three diagnostic peaks, at least the first strong diagnostic peak, were identified at multiple sites and were effective at differentiating case/control samples at all sites. In addition, the likelihood of separate sites identifying the same peaks has been disputed *(38, 39)*. We have not revealed the identities (i.e., the *m/z* values) of the diagnostic peaks used in this study because the ongoing studies (phase II) require the sites to remain blind to the identities of these peaks. Although concern for lack of "analytical sensitivity" and whether the diagnostic peaks are from "high-abundant molecules" was raised, these issues alone become more academic if the assay is clinically useful and reproducible. However, if, as we have described, the larger, more prominent peaks are the most reproducible, then indeed this process is limited at the level of effective analytical sensitivity. Specifically, improvements in the sensitivity that are reflected in stronger *m/z* peak S/N will improve the robustness of measurements; consequently, upfront sample depletion and concentration steps may be needed. Ultimately, what was experimentally called for was a "published report that similar data can be obtained by using different batches of SELDI chips, different technologists, or by using the same conditions at a later time". Our current EPSIC study provides this information. We have demonstrated that under the strict operating procedures that we have described, we are able to achieve across-laboratory reproducibility of SELDI-TOF-MS analysis. We would therefore suggest that the next important step in the evaluation of this assay approach is a demonstration of the robustness of clinical classification as is slated for our phase II design.

In conclusion, we show that, if the protocols described here are adhered to, SELDI-TOF-MS profiling can provide a reproducible diagnostic assay platform, providing that the measured events are separable *m/z* peak values. We are currently examining the robustness of the assay across patient sample sets. This phase II analysis will address the ability of the current combination of SELDI-TOF-MS assay and decision algorithm to perform as a robust diagnostic tool for the detection of prostate cancer.

We would like to acknowledge Jim Jones of Ciphergen Biosystems for technical support and dedication to this project. This work was supported by the Early Detection Research Network, National Cancer Institute (U01CA084986) and Grants CA86402 (to I.T.), CA84968 (to W.L.B.), CA86368 (to Z.F.), CA86359 (to W.E.G.), CA85067 (to O.J.S.), and CA86(32)3 (to A.P.).

References

1. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, et al. The case for early detection. Nat Rev Cancer 2003;3:243–52.
2. Gloeckler-Ries LA, Reichman ME, Lewis DR, Hankey BF, Edwards BK. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. Oncologist 2003;8:541–52.
3. Jemal A, Tiwari RC, Murray T, Ghafoor A, Samuels A, Ward E, et al. Cancer statistics, 2004. CA Cancer J Clin 2004;54:8–29.
4. Tiwari RC, Ghosh K, Jemal A, Hachey M, Ward E, Thun MJ, et al. A new method of predicting US and state-level cancer mortality counts for the current calendar year. CA Cancer J Clin 2004;54:30–40.
5. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer 2002;359:572–7.
6. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 2002;62:3609–14.
7. Vlahou A, Laronga C, Wilson L, Gregory B, Fournier K, McGaughey D, et al. A novel approach toward development of a rapid blood test for breast cancer. Clin Breast Cancer 2003;4:203–9.
8. 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.
9. Banez LL, Prasanna P, Sun L, Ali A, Zou Z, Adam BL, et al. Diagnostic potential of serum proteomic patterns in prostate cancer. J Urol 2003;170:442–6.
10. Cazares LH, Adam BL, Ward MD, Nasim S, Schellhammer PF, Semmes OJ, et al. Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry. Clin Cancer Res 2002;8:2541–52.
11. Koopmann J, Zhang Z, White N, Rosenzweig J, Fedarko N, Jagannath S, et al. Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. Clin Cancer Res 2004;10:860–8.
12. Kozak KR, Amneus MW, Pusey SM, Su F, Luong MN, Luong SA, et al. Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. Proc Natl Acad Sci U S A 2003;100:12343–8.
13. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 2002;48:1296–304.
14. Paweletz CP, Trock B, Pennanen M, Tsangaris T, Magnant C, Liotta LA, et al. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. Dis Markers 2001;17:301–7.
15. Petricoin EF 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, et al. Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2002;94:1576–8.
16. Poon TC, Yip TT, Chan AT, Yip C, Yip V, Mok TS, et al. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem 2003;49:752–60.
17. Rosty C, Christa L, Kuzdzal S, Baldwin WM, Zahurak ML, Carnot F, et al. Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology. Cancer Res 2002;62:1868–75.
18. Vlahou A, Schellhammer PF, Mendrinos S, Patel K, Kondylis FI, Gong L, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001;158:1491–502.
19. Wadsworth JT, Somers KD, Cazares LH, Malik G, Adam BL, Stack BC Jr, et al. Serum protein profiles to identify head and neck cancer. Clin Cancer Res 2004;10:1625–32.
20. Wadsworth JT, Somers KD, Stack BC Jr, Cazares L, Malik G, Adam BL, et al. Identification of patients with head and neck cancer using serum protein profiles. Arch Otolaryngol Head Neck Surg 2004;130:98–104.
21. Won Y, Song HJ, Kang TW, Kim JJ, Han BD, Lee SW. Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. Proteomics 2003;3:2310–6.

22. Wulfkuhle JD, McLean KC, Paweletz CP, Sgroi DC, Trock BJ, Steeg PS, et al. New approaches to proteomic analysis of breast cancer. Proteomics 2001;1:1205–15.

23. Xiao X, Liu D, Tang Y, Guo F, Xia L, Liu J, et al. Development of proteomic patterns for detecting lung cancer. Dis Markers 2003; 19:33–9.

24. Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS. Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. Lung Cancer 2003;40:267–79.

25. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level, < or = 4.0 ng per milliliter. N Engl J Med 2004;22:2239–46.

26. Check E. Proteomics and cancer: running before we can walk? Nature 2004;429:496–7.

27. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 2004;96:353–6.

28. Diamandis EP. Re: diagnostic potential of serum proteomic patterns in prostate cancer. J Urol 2004;171:1244–5 [author reply 124–5-64].

29. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4: 24.

30. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20:777–85.

31. Gao CL, Rawal SK, Sun L, Ali A, Connelly RR, Banez LL, et al. Diagnostic potential of prostate-specific antigen expressing epithelial cells in blood of prostate cancer patients. Clin Cancer Res 2003;9:2545–50.

32. Grizzle WE, Adam BL, Bigbee WL, Conrads TP, Carroll C, Feng Z, et al. Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer. Dis Markers 2003;19:185–95.

33. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL Jr, Qu Y, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics 2003;4:449–63.

34. Yasui Y, McLerran D, Adam BL, Winget M, Thornquist M, Feng Z. An automated peak-identification/calibration procedure for high-dimensional protein measures from mass spectrometers. J Biomed Biotechnol 2003;2003:242–8.

35. Sullivan Pepe M, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer [Review]. J Natl Cancer Inst 2001;93:1054–61.

36. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309–14.

37. Grizzle WE, Semmes OJ, Basler J, Izbicka E, Feng Z, Kagan J, et al. The early detection research network surface-enhanced laser desorption and ionization prostate cancer detection study: a study in biomarker validation in genitourinary oncology. Urol Oncol 2004;22:337–43.

38. Grizzle WE, Meleth S. Clarification in the Point/Counterpoint discussion related to surface-enhanced laser desorption/ionization time-of-flight mass spectrometric identification of patients with adenocarcinomas of the prostate [Letter]. Clin Chem 2004; 50:1475–6.

39. Petricoin EF. Liotta LA. Proteomic pattern complexity reveals a rich and uncharted continent of biomarkers [Reply]. Clin Chem 2004; 50:1476–7.