REVIEW

High-resolution serum proteomic features for ovarian cancer detection

T P Conrads[1], *V A Fusaro*[2,3], *S Ross*[2], *D Johann*[2,3], *V Rajapakse*[2,3], *B A Hitt*[4], *S M Steinberg*[5], *E C Kohn*[3], *D A Fishman*[6], *G Whiteley*[7], *J C Barrett*[8], *L A Liotta*[3], *E F Petricoin III*[2] and *T D Veenstra*[1]

[1]National Cancer Institute Biomedical Proteomics Program, Laboratory of Proteomics and Analytical Technologies, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, MD 21702, USA
[2]Food and Drug Administration-National Cancer Institute Clinical Proteomics Program, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD 20892, USA
[3]Laboratory of Pathology, National Cancer Institute, Center for Cancer Research, Bethesda, MD 20892, USA
[4]Correlogic Systems, Inc., Bethesda, MD 20892, USA
[5]Biostatistics and Data Management Section, National Cancer Institute, Center for Cancer Research, Bethesda, MD 20892, USA
[6]National Ovarian Cancer Early Detection Program, Northwestern University, Chicago, IL 60611, USA
[7]Serum Proteomic Patterns Clinical Reference Laboratory, National Cancer Institute at Frederick, SAIC-Frederick, Inc., Frederick, MD 21702, USA
[8]Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA

(Requests for offprints should be addressed to T D Veenstra, SAIC-Frederick, National Cancer Institute at Frederick, P.O. Box B, Frederick, MD 21702, USA; Email: veenstra@ncifcrf.gov)

Abstract

Serum proteomic pattern diagnostics is an emerging paradigm employing low-resolution mass spectrometry (MS) to generate a set of biomarker classifiers. In the present study, we utilized a well-controlled ovarian cancer serum study set to compare the sensitivity and specificity of serum proteomic diagnostic patterns acquired using a high-resolution versus a low-resolution MS platform. In blinded testing sets, the high-resolution mass spectral data contained multiple diagnostic signatures that were superior to the low-resolution spectra in terms of sensitivity and specificity ($P < 0.00001$) throughout the range of modeling conditions. Four mass spectral feature set patterns acquired from data obtained exclusively with the high-resolution mass spectrometer were 100% specific and sensitive in their diagnosis of serum samples as being acquired from either unaffected patients or those suffering from ovarian cancer. Important to the future of proteomic pattern diagnostics is the ability to recognize inferior spectra statistically, so that those resulting from a specific process error are recognized prior to their potentially incorrect (and damaging) diagnosis. To meet this need, we have developed a series of quality-assurance and in-process control procedures to (a) globally evaluate sources of sample variability, (b) identify outlying mass spectra, and (c) develop quality-control release specifications. From these quality-assurance and control (QA/QC) specifications, we identified 32 mass spectra out of the total 248 that showed statistically significant differences from the norm. Hence, 216 of the initial 248 high-resolution mass spectra were determined to be of high quality and were remodeled by pattern-recognition analysis. Again, we obtained four mass spectral feature set patterns that also exhibited 100% sensitivity and specificity in blinded validation tests (68/68 cancer: including 18/18 stage I, and 43/43 healthy). We conclude that (a) the use of high-resolution MS yields superior classification patterns as compared with those obtained with lower resolution instrumentation; (b) although the process error that we discovered did not have a deleterious impact on the present results obtained from proteomic pattern analysis, the major source of spectral variability emanated from mass spectral acquisition, and not bias at the clinical collection site; (c) this variability can be reduced and monitored through the use of QA/QC statistical procedures; (d) multiple and distinct proteomic patterns, comprising low molecular weight biomarkers, detected by high-resolution MS achieve accuracies surpassing individual biomarkers, warranting validation in a large clinical study.

Endocrine-Related Cancer (2004) **11** 163–178

1351-0088/04/011–163 © 2004 Society for Endocrinology *Printed in Great Britain*

Online version via http://www.endocrinology.org

Introduction

Ovarian cancer is the most lethal gynecologic malignancy and is the fifth most common cause of cancer-related death in women. The American Cancer Society estimated that there would be 25 400 new cases of ovarian cancer and 14 300 deaths in 2003 (www.cancer.org). Since the 1960s, almost 80% of women with epithelial ovarian cancer are diagnosed when the disease has spread to the upper abdomen (stage III) or beyond (stage IV) (Niloff *et al*. 1984, Menon *et al*. 2000, Cohen *et al*. 2001, Ozols 2001). Unfortunately, the 5-year survival rate for those women is approximately 15%, whereas the 5-year survival when detected at early stage (I) approaches 90% (Niloff *et al*. 1984, Menon *et al*. 2000, Cohen *et al*. 2001, Ozols 2001). Therefore, the diagnosis of early stage ovarian cancer would significantly decrease the morbidity and mortality rate from this disease.

The most widely used diagnostic biomarker for ovarian cancer is cancer antigen 125 (CA 125), as detected by the monoclonal antibody OC 125 (Zurawski *et al*. 1988). Alhough 80% of women with advanced-stage ovarian cancer possess elevated levels of CA 125, it is elevated in only 50–60% of women with stage I disease (Niloff *et al*. 1984, Menon *et al*. 2000, Cohen *et al*. 2001, Ozols 2001), with a positive-predictive value of 10%. Moreover, CA 125 can be elevated in many other nongynecologic and benign conditions, such as pregnancy, endometriosis, and colon and pancreatic cancers. A combined strategy of CA 125 determination with ultrasonography increases the positive-predictive value to approximately 20% (Cohen *et al*. 2001). Consequently, there is an urgent need to develop detection methods to improve the sensitivity and specificity of early-stage ovarian cancer detection.

Several laboratories have demonstrated the feasibility of using serum-based proteomic pattern analysis by mass spectrometry (MS) for the diagnosis of ovarian (Petricoin et al. 2002*a*), breast (Li *et al*. 2002) and prostate (Adam *et al*. 2002, Petricoin *et al*. 2002*a*) cancer. Unlike previous biomarker discovery efforts that attempt to identify a single disease biomarker candidate, proteomic pattern analysis utilizes high-dimensional data such as data resulting from MS analyses. This method attempts, without bias, to identify patterns of mass spectral features comprising peptide (or other) ions within mass spectra as the diagnostic itself. Utilizing serum, for example, mass spectra generated from a training set of samples are analyzed by pattern-recognition algorithms to identify diagnostic signature patterns comprising a subset of key mass-to-charge ratio (m/z) species and their relative intensities. Mass spectra from unknown samples are subsequently classified by likeness to the pattern found in the serum mass spectra used in the training set. The number of key m/z species whose combined relative intensities define the pattern represents a very small subset of the entire number of species present in any given serum mass spectrum.

Petricoin *et al*. (2002*a*) have recently demonstrated that serum proteomic patterns from low-resolution MS data can distinguish neoplastic from nonneoplastic disease within the ovary. A key aspect of their study was the application of a pattern-recognition tool that employs an unsupervised system (self-organizing-type cluster mapping) as a fitness test for a supervised system (a genetic algorithm). With their approach, a training set comprising mass spectra from serum derived from either unaffected women or women with ovarian cancer is employed so that the most fit combination of relative, normalized ion-intensity features defined at precise m/z values plotted in n-space can reliably distinguish the cohorts used in training. While this approach first demonstrated the feasibility of a proteomic pattern-based diagnostic test, translating this approach to a routine clinical diagnostic test remains a daunting challenge. Specifically troubling to this translation is the fact that the original MS used for the early feasibility studies, the ProteinChip Biomarker System-II (PBS-II), a low-resolution time-of-flight (TOF) MS, is a research-grade platform not designed for routine clinical use. While it is reproducible within runs and small intervals of time, we observed that week-to-week and machine-to-machine variability was unacceptable as a general clinical method in its current state in our hands. Moreover, the resolution of the original MS was not sufficient to resolve species close in m/z but rather gave rise to coalesced features, severely compromising unique feature selection in the diagnostic pattern discovery. The need exists to extend these observations to MS with performance characteristics aligned with routine clinical use: high-resolution and reduced day-to-day mass drift as can be accomplished by decoupling the source from the mass analyzer, the basis of the design of the hybrid quadrupole time-of-flight (QqTOF) MS, which is employed in this study as described below.

The first objective of the present study was to compare mass spectra from a high-resolution and a low-resolution mass spectrometer, using sera obtained from a large, well-controlled ovarian cancer screening trial applied and analyzed on the same SELDI ProteinChip arrays (Hutchens *et al*. 1993). While we hypothesize that higher resolution mass spectra will generate more distinguishable sets of diagnostic features, the increased complexity and dimensionality of data may actually reduce the likelihood of fruitful pattern discovery. Moreover, we now describe spectral quality-assurance and control (QA/QC) methods whereby mass spectra are analyzed for overall intensity and complexity prior to pattern-recognition analysis to reduce experimental and sample variability introduced by the process instead of the disease. In a clinical setting where a pattern test may be eventually employed as a diagnostic, it will be crucial to determine overall spectral quality and develop spectral QA/QC release specifications such that variances introduced into the process can be evaluated and monitored. Detailed procedures to detect day-to-day, lot-to-lot and machine-to-machine variances arising from sample handling, storage and shipping conditions, as well as fluctuations in performance by the MS, must be developed and implemented. Therefore, the second objective was to develop and implement a series of QA/QC procedures to statistically evaluate mass spectra. The spectra that passed this QA/QC procedure were reanalyzed to identify sets of features with the best overall specificity and sensitivity. We report here the use of serum proteomic pattern analysis for the generation of multiple, highly accurate models obtained with a QqTOF MS for an improved early diagnosis of ovarian cancer. In addition, we propose QA/QC methods to evaluate spectra prior to bioinformatic analysis. This evaluation allows specific procedural errors that may occur during the analysis to be recognized, and it can prevent the use of spectra that can lead to an incorrect diagnosis.

Materials and methods

Serum samples

Serum samples were obtained from the National Ovarian Cancer Early Detection Program (NOCEDP) and gynecologic oncology clinic at Northwestern University (Chicago, IL, USA). Specimens from women enrolled in the NOCEDP who had no evidence of any cancer for 5 years were evaluated as being from healthy women. Similarly, only preoperative specimens were used from women who were surgically staged and found to have epithelial ovarian carcinoma. A total of 248 samples were prepared with a Biomek 2000 robotic liquid handler (Beckman Coulter, Inc., Palo Alto, CA, USA). All analyses used ProteinChip weak cation exchange interaction chips (WCX2, Ciphergen Biosystems, Inc., Fremont, CA, USA). A control reference sample was randomly applied to one spot on each protein array as a quality control for overall process integrity, sample preparation and mass spectrometer function. The control sample, SRM 1951A, which comprises pooled normal human sera, was provided by the National Institute of Standards and Technology (Gaithersburg, MD, USA).

Sample preparation

WCX2 ProteinChip arrays were processed in parallel on a Biomek Laboratory workstation (Beckman-Coulter) modified to make use of a ProteinChip array bioprocessor (Ciphergen Biosystems). The bioprocessor holds 12 ProteinChips, each having eight chromatographic ‘spots’, allowing 96 samples to be processed in parallel. A volume of 100 µl of 10 mM HCl was applied to the WCX2 protein arrays and allowed to incubate for 5 min. The HCl was aspirated and discarded, and 100 µl distilled, deionized water (ddH$_2$O) were applied and allowed to incubate for 1 min. The ddH$_2$O was aspirated, discarded, and reapplied for another minute. A volume of 100 µl of 10 mM NH$_4$HCO$_3$ with 0.1% Triton X-100 was applied to the surface and allowed to incubate for 5 min, after which the solution was aspirated and discarded. A second application of 100 µl of 10 mM NH$_4$HCO$_3$ with 0.1% Triton X-100 was applied and allowed to incubate for 5 min, after which the ProteinChip array bait surfaces were aspirated. A volume of 5 µl of raw, undiluted serum was applied to each ProteinChip WCX2 bait surface and allowed to incubate for 55 min. Each ProteinChip array was washed three times with Dulbecco’s phosphate-buffered saline and ddH$_2$O. For each wash, 150 µl of either phosphate-buffered saline or ddH$_2$O were sequentially dispensed, mixed by aspirating, and dispensed for a total of 10 times in the bioprocessor, after which the solution was aspirated to waste. This wash process was repeated for a total of six washes per ProteinChip array bait surface. The ProteinChip array bait surfaces were vacuum dried to prevent cross-contamination when the bioprocessor gasket was removed. After removal of the bioprocessor gasket, 1.0 µl of a 30% solution of α-cyano-5-hydroxycinnamic acid in 50% (v/v) acetonitrile and 0.5% (v/v) trifluoroacetic acid was applied to each spot on the ProteinChip array twice, allowing the applied solution to dry between applications with a liquid robotic handling station Genesis Freedom 200 (TECAN, Research Triangle Park, NC, USA).

PBS-II TOF MS analysis

ProteinChip arrays were placed in the Protein Biological System II time-of-flight (TOF) mass spectrometer ((PBS-II, Ciphergen Biosystems), and mass spectra were recorded on the following settings: 195 laser shots/spectrum collected in positive ionization mode, laser intensity 220, detector sensitivity 5, detector voltage 1850 V, and time-lag focus of 6000 m/z. The PBS-II TOF MS was externally calibrated with the ‘All-In-One’ peptide mass standard (Ciphergen Biosystems).

QqTOF MS analysis

ProteinChip arrays were analyzed with a hybrid quadrupole time-of-flight mass spectrometer (QSTAR pulsar *I*, Applied Biosystems, Inc., Framingham, MA, USA) fitted with a ProteinChip array interface (Ciphergen Biosystems). Samples were ionized with a 337 nm pulsed nitrogen laser (ThermoLaser Sciences model VSL-337-ND-S, Waltham, MA, USA) operating at 30 Hz. Approximately 20 mTorr of nitrogen gas was used for collisional ion cooling. Each spectrum represents 100 multichannel averaged scans (1.667 min acquisition/spectrum). The mass spectrometer was externally calibrated with a mixture of known peptides.

Proteomic pattern analysis

Proteomic pattern analysis was performed by exporting the raw data file generated from the PBS-II into tab-delimited files possessing approximately 15 000 data points. The QqTOF mass spectra were similarly exported into a tab-delimited format possessing approximately 350 000 data points per spectrum. The high-resolution spectra were binned using a 400 parts per million (ppm) function to produce data files that possess identical m/z values (for example, the m/z bin sizes scale linearly from a bin width of 0.28 at m/z 700 to 4.75 at m/z 12 000). This binning condenses the number of data points from 350 000 to exactly 7084 points per spectrum. The conservative 400 ppm binning function was based on the value obtained by 10 times the routine mass accuracy of the QqTOF with external calibration (40–50 ppm). The mass spectra were randomly segregated into equal groups for training, and testing. The models were built on the training set, using ProteomeQuest (Correlogic Systems, Inc., Bethesda, MD, USA), and tested in blinded sample sets. The m/z values in the models that were generated by the high-resolution instrument are based on the binned data, and not the actual m/z values from the raw mass spectra.

The Proteome Quest software itself implements a pattern discovery algorithm combining elements from genetic algorithms (Holland 1994) and self-organizing adaptive pattern-recognition systems (Kohonen 1990). Genetic algorithms organize and analyze complex data sets as if they were information comprising individual elements that can be manipulated through a computer-driven analog of a natural selection process. Self-organizing systems cluster data patterns into similar groups. Adaptive systems recognize novel events and track rare instances. The genetic algorithm component of the analysis begins with the random generation of a population of 1500 subsets of combinations of features in the serum mass spectra. The choice of this number was based on adequate coverage of the data, with a heuristic that no value can be duplicated within each of the 1500 feature subsets. Each feature subset in the population specifies the identities of the exact m/z values in each serum mass spectrum, but not their relative amplitude. The number of features in the subset ranges from 5 to 20. Data normalization is an important element of pattern recognition, as bias introduced by ProteinChip quality, MS performance and operator variance can affect the overall spectral quality. Since the present MS technique is not inherently quantitative, scalar MS peak intensity changes may be apparent, yet the overall pattern may not change. For this study, MS data were normalized by linearly scaling each m/z value, V, within any randomly generated pattern subset between the largest and the smallest values within that subset, so that $0 \leq NV \leq 1$. In this way, differences in spectral quality that may emanate from biases, such as ProteinChip variance, and not from the inherent disease process itself, can be minimized. The spectra are normalized according to the following formula:

$$NV = (V - \text{Min})/(\text{Max} - \text{Min})$$

where NV is the normalized m/z value, V is the intensity value for the specific randomly chosen m/z bin, Min is the intensity of the smallest intensity value of any of the m/z bins within the randomly selected feature set and Max is the maximum intensity of the m/z bin within the randomly selected feature set. This equation linearly normalizes the peak intensities in the feature set so as to fall within the range of 0 to 1. Prior to analysis, the data are randomly divided into training and testing data sets. The training data set is further divided into and labeled as diseased or unaffected according to known clinical diagnosis.

Each of the randomly selected 1500 subset feature sets was subjected to a fitness test. The fitness test in these analyses is the ability of the combined m/z amplitude values of any candidate feature set to specify a lead cluster map that generates homogeneous clusters containing only mass spectra of diseased subjects or unaffected subjects used in the training sets. The lead cluster map is a self-organizing, adaptive pattern-recognition algorithm that uses Euclidean distance to group vectors of data. The map begins as empty n-dimensional space where n is the number of m/z features in the data vector. The optimal discriminatory pattern is identified by finding the best combination of m/z bins whose normalized feature set intensity values in n-dimensional space creates a unique identifier or cluster of identifiers. Any given training sample is compared for its proximity to previously defined clusters of diseased and unaffected subjects in n-space. If an n-dimensional identifier vector from a subject in the training group falls within the decision boundary of an existing cluster, the subject is classified as belonging to that group. For these studies, the decision boundary is defined as 10% of the maximum distance allowed in the space. The population that lies within this boundary corresponds to a 90% pattern match. If the data vector does not fall within the 90% decision boundary of any existing cluster in the model, it is used to establish a new cluster and is identified as a new observation. The process is repeated once for each vector in the collection of training data.

Those subpopulation feature sets that best discriminate the training set are more likely to survive the culling of the population to the original population size, such as 1500, and contribute to the next generation of fit candidate feature sets. The progeny of the most-fit feature sets are generated through crossover and mutation of the 5–20 specific *m*/*z* bin values within each subset. Each subset is evaluated for its ability to distinguish accurately the two training set populations. As a result, each successive population of feature subsets is, on average, more fit than its predecessor feature sets. To ensure that the algorithms do not trend to less than near optimal decision points, a 'mutation' rate is built into the process such that 0.2% of the *m*/*z* bin values are randomly rechosen. Crossover operations are of the single point type and are randomly selected in each mating. For example, if there are five *m*/*z* bin values, there can be four crossover points. The genetic algorithm iterates for at least 250 generations or until a lead cluster map that homogeneously segregates diseased serum mass spectra from unaffected is generated. The lead cluster map that best separates diseased from unaffected serum mass spectra is deployed for validation in blinded test sets.

Completely blinded test data, not used during the training process, were analyzed in the following steps. The data were normalized as described, and the normalized relative amplitudes of the test sample spectra at the *n* defined *m*/*z* values were used to fix a point in *n*-dimensional space. The Euclidean distance vector was then calculated between this point and the center of all clusters (both cancer and unaffected) formed by the training set. If the unknown test vector fell inside the 90% boundary surrounding any centroid, it was classified as a member of that cluster and given a probability score based on its proximity to the theoretical center of the cluster and the number of records within that cluster. If no match was obtained, it was scored as a 'new cluster'. The results from the testing set of data were used for determination of sensitivity, specificity and positive predictive value of the patterns.

Although it is impossible to visualize plots of points with more than three coordinates, Pythagorean based formulas adapt quite satisfactorily to points in higher dimensions, including the formula for the distance between two points. For example, the distance between two points in five dimensions, (a1, b1, c1, d1, e1) and (a2, b2, c2, d2, e2), is calculated as follows:

$$\text{Distance} = \surd(\text{a1} - \text{a2})^2 + (\text{b1} - \text{b2})^2 + (\text{c1} - \text{c2})^2 + (\text{d1} - \text{d2})^2 + (\text{e1} - \text{e2})^2],$$

as distance always equals the square root of the sum of squared differences between coordinates. Hence, the number of squared terms equals the number of members within the selected subset pattern.

Spectral quality control and quality assurance

The total ion current (TIC) of the raw and binned mass spectral data was plotted, average/mean and standard deviation of amplitude were calculated, chi-square and *t*-test analysis of each *m*/*z* or bin value, and quartile plotting measured with JMP (SAS Institute, Cary, NC, USA) software as well as procedures developed in-house. Process measures were checked by analyzing the statistical plots of the serum reference standard (SRM-015A, National Institute of Standards and Technology) that was applied at random on each ProteinChip at different spot locations.

Statistical analysis

The exact Cochran–Armitage test for trend (Agresti 1990) was used to obtain the statistical significance of the differences of the distributions of sensitivity and specificity for the QqTOF and PBS-II models generated according to the procedures detailed above. All *P*-values are two-tailed, as indicated by P_2.

Results

Comparative analysis of serum samples by low- and high-resolution MS

A total of 248 serum samples were provided from the NOCEDP and gynecologic oncology clinic at Northwestern University. The samples were processed and their proteomic patterns acquired from the same ProteinChip arrays by both a PBS-II and a QqTOF MS fitted with a ProteinChip interface (PCI-1000). The region of the sample queried by the laser on the different instruments does not overlap, thus affording the ability to have each mass spectrometer analyze the same spot on the same ProteinChip. While the mass spectra acquired from both instruments are qualitatively similar, the higher resolution afforded by the QqTOF MS is readily apparent (Fig. 1). This increased resolution allows species with similar *m*/*z* values that are unresolved by the PBS-II TOF MS to be resolved in the QqTOF mass spectrum. Indeed, simulations demonstrate the ability of the QqTOF MS (routine

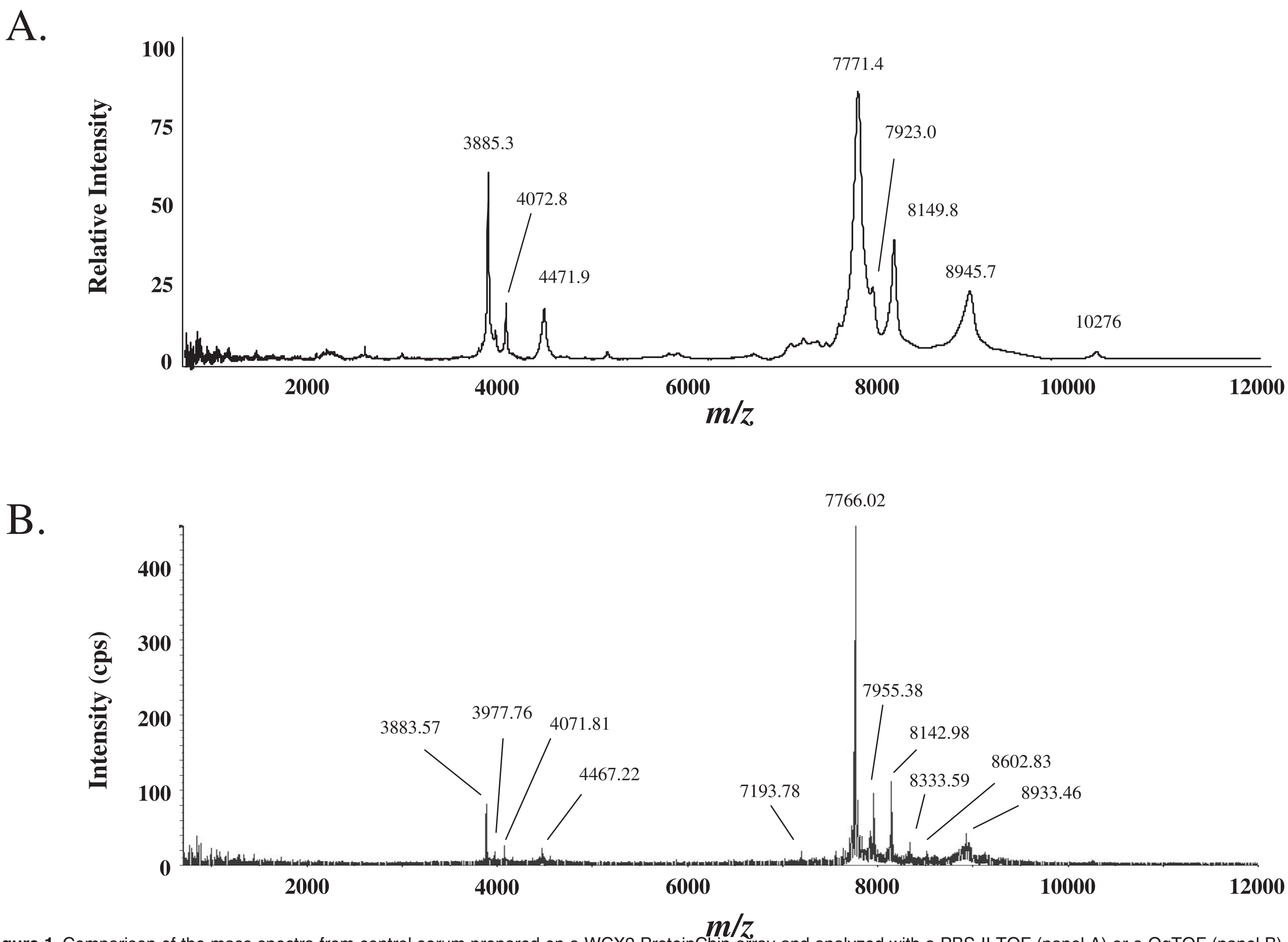


Figure 1 Comparison of the mass spectra from control serum prepared on a WCX2 ProteinChip array and analyzed with a PBS-II TOF (panel A) or a QqTOF (panel B) mass spectrometer.

resolution > 8000) to resolve completely species differing in m/z of only 0.375 (for example, at m/z 3000), whereas complete resolution of species with the PBS-II TOF MS (routine resolution of ~150) is possible only for species that differ by m/z of 20 (simulation not shown).

The mass spectra were analyzed with the ProteomeQuest bioinformatics tool employing ASCII files consisting of m/z and intensity values derived from either the PBS-II TOF or the QqTOF mass spectra as the input. The mass spectral data acquired using the QqTOF MS were binned to define precisely the number of features in each spectrum to 7084, with each feature comprising a binned m/z and amplitude value. The algorithm examines the data to find a set of features at precise m/z values whose combined, normalized relative intensity values in n-space best segregate the data derived from the training set. Mass spectra acquired on the QqTOF and the PBS-II TOF instruments from the same sample sets were restricted to the m/z range from 700 to 12 000 for direct comparison between the two platforms. The entire set of spectra acquired from the serum samples was divided into three data sets: (a) a training set that is used to discover the hidden diagnostic patterns, (b) a testing set, and (c) a validation set. Only the normalized intensities of the key subset of m/z values identified using the training set were used to classify the testing and validation sets, and the algorithm had not previously 'seen' the spectra in the testing and validation sets. The training set comprised serum from 28 healthy women and 49 women with epithelial ovarian cancer. The training and testing set mass spectra were analyzed by the bioinformatic algorithm to generate a series of models under the following set of modeling parameters: (a) a similarity space of 85%, 90%, or 95% likeness for cluster classification; (b) a feature set size of 5, 10 or 15 random m/z values whose combined intensities comprise each pattern; and (c) a learning rate of 0.1%, 0.2% or 0.3% for pattern generation by the genetic algorithm. Four sets of randomly generated models for each of the 27 permutations were derived and queried with the same test set. Sensitivity- and specificity-blind testing and validation results for each of the 108 models (four rounds of training for each of the 27 permutations) were generated (Fig. 2). These results demonstrate that the serum mass spectra from the QqTOF consistently outperformed mass spectra obtained on the lower resolution MS in terms of sensitivity ($P_2 < 0.00001$), where P_2 denotes a two-tailed Cochran–Armitage test for trend (Agresti 1990)) and specificity ($P_2 < 0.00001$) throughout the range of modeling conditions, for this specific set of clinical sera.

Evaluation of the models diagnostic for ovarian cancer

The ability to generate the best performing models for testing and validation was statistically evaluated, as multiple models were generated and ranked using the entire range of the aforementioned modeling parameters. Models from the training set were assessed with a blinded testing set consisting of spectra acquired from serum obtained from 31 unaffected and 63 ovarian cancer-affected individuals. For further validation of the ability to diagnose ovarian cancer, a set of spectra acquired from blinded samples comprising 37 normal and 40 ovarian cancer serum mass spectra were tested against the models found in training, as previously discussed. The results (Fig. 2) clearly show the ability of the mass spectra from the higher resolution QqTOF MS to generate statistically superior models over the lower resolution PBS-II TOF mass spectra in both sensitivity ($P_2 < 0.0001$) and specificity ($P_2 < 3 \times 10^{-19}$) in the testing phase, as well as sensitivity ($P_2 < 9 \times 10^{-9}$) and specificity ($P_2 < 6 \times 10^{-6}$) in the blinded validation phase.

Four models were found that were both 100% sensitive and specific in their ability to discriminate correctly mass spectra acquired from serum samples obtained from unaffected women from those suffering from ovarian cancer (Fig. 3). All of these models were obtained with data acquired using the QqTOF MS, as *no* models generated using the PBS-II TOF MS data were both 100% sensitive and specific. Examination of the key m/z features that comprise the four best performing patterns reveals certain features (that is, contained within m/z bins 7060.121, 8605.678 and 8706.065) that are consistently present as classifiers in the models (Fig. 3). Although the proteomic patterns generated from both healthy and cancer patients with the QqTOF MS are quite similar (Fig. 4), careful inspection of the raw mass spectra reveals that peaks within the binned m/z values 7060.121 and 8605.678 are indeed differentially abundant in a selection of the serum samples obtained from ovarian cancer patients as compared with unaffected individuals (Fig. 4, insets). These results indicate that these MS peaks originate from species that may be consistent indicators of the presence of ovarian cancer. However, the ability to distinguish serum from an unaffected individual or one with ovarian cancer by a single serum proteomic m/z feature alone is not possible across the entire serum study set. While a single key m/z species is insufficient to distinguish globally all of the unaffected and ovarian cancer patients, the combined peak intensities of key ions, taken together, do allow the two data sets to be completely distinguished.

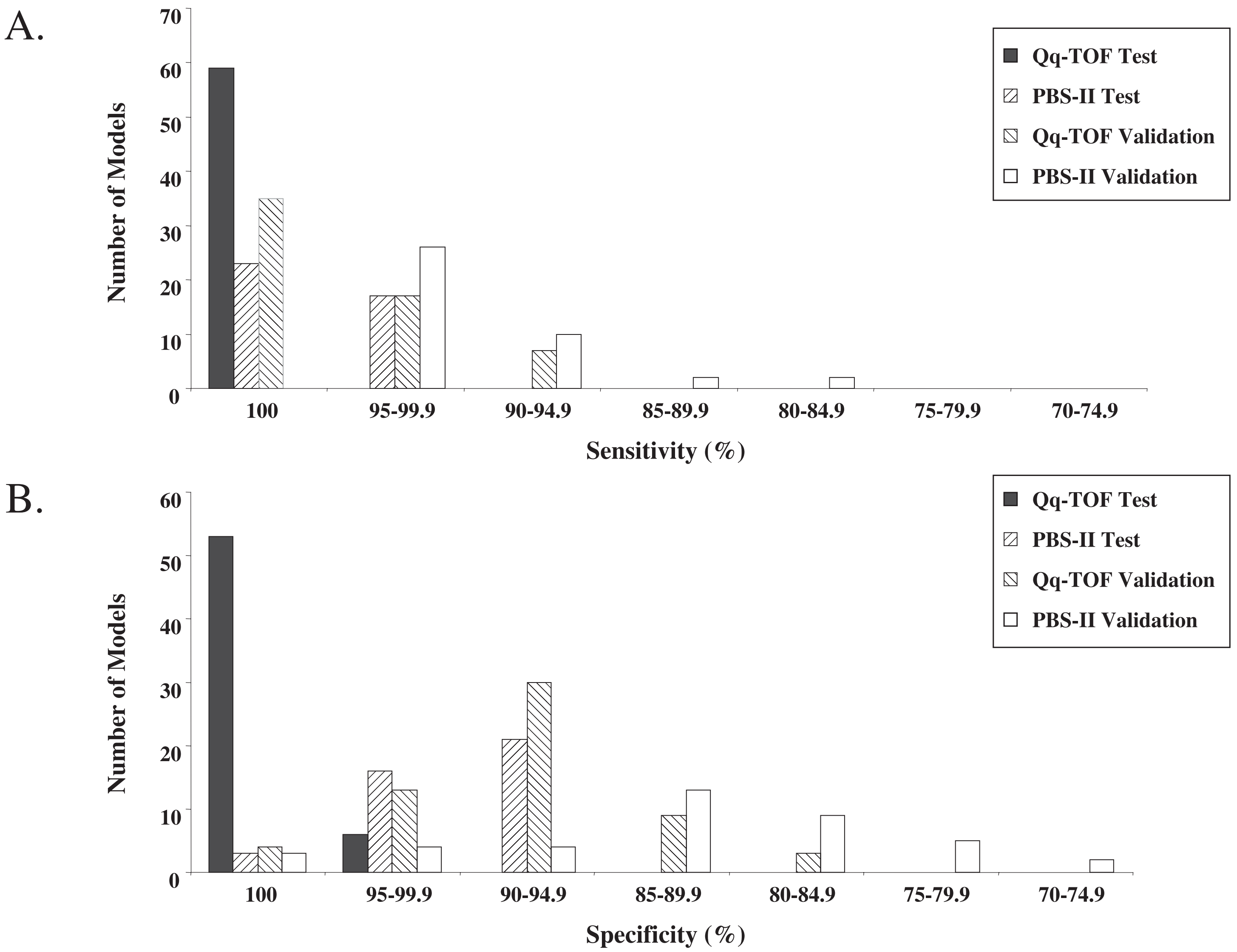


Figure 2 Histograms representing the testing and blinded validation results of sensitivity (A) and specificity (B) of the 108 models for MS data acquired on either a QqTOF or a PBS-II TOF mass spectrometer.

Sensitivity: 100% (103/103 Cancer; 22/22 stage I)
Specificity: 100% (67/67 Healthy)

1	2	3	4
1276.861	818.480	1144.796	1001.654
2374.244	6352.723	4260.403	1255.593
4292.900	6548.771	7046.018	4377.853
7060.121	7060.121	8602.237	6004.416
8605.678	7096.922	8664.385	7060.121
8706.065	8540.536		7202.716
9870.937	8605.678		8605.678
	8706.065		8709.548
			9367.113

Figure 3 Four distinct models that generated 100% accuracy in testing and validation and the key *m/z* values whose combined intensities discriminated the sets. Common *m/z* features recurring between the models are highlighted.

Application of quality-control analysis

While the results provided by the data acquired with the QqTOF MS were of superior diagnostic value, if proteomic pattern profiling is to be a viable clinical tool, it will be critical to develop a set of statistical tools enabling the recognition of spectra quality prior to their bioinformatic analysis. This statistical QA/QC procedure is critical to prevent incorrect diagnosis resulting from a process error and to recognize errors that may occur in the overall procedure. To meet this need, the 248 mass spectra acquired by the QqTOF MS were analyzed with a wide variety of statistical tools to evaluate the spectral quality (that is, record count and mean amplitude) and statistical variances greater than the population norm. Mean spectral amplitude and the file record count (that is, the total number of data points within a mass spectrum) were selected as global parameters for statistical analysis. Using these parameters, the mass spectra from serum of unaffected individuals and cancer patients were tracked over several days of analysis and over different lots of ProteinChips (Figs 5 and 6). To distinguish variability related to the clinical sample, the sampling process, or the MS instrument, control reference samples were randomly analyzed in parallel on each ProteinChip (Fig. 5). The total record count of each mass spectrum recorded of the reference sample is plotted in Fig. 5A. The results show a trend toward lower total record count as the analysis progressed, indicating a potential process error. The mean amplitude of each mass spectrum was plotted against ProteinChip lot number (Fig. 5B) and day on which each mass spectrum was acquired (Fig. 5C). Plots of the total record count (Fig. 6A), the ProteinChip lot-dependent mean amplitude (Fig. 6B) and the acquisition time-dependent mean amplitude (Fig. 6C) for the actual serum samples showed trends reflective of those observed for the reference sample mass spectra. The results of these QA/QC measurements indicated 32 spectra that were of lesser quality based on total record count, amplitude mean and standard deviation error (Fig. 6, marked by asterisks). These mass spectra were all generated at the end of the experimental run, suggesting that a deviation in the process had occurred. This process variance was determined to be due not to ProteinChip lot-to-lot variation, but to a failing grid in the accelerator region of the QqTOF MS. Intriguingly, the spectral QA/QC procedures developed here were able to indicate this failure early on, demonstrating the functional utility and value of using release-specification and in-process controls. Importantly, the total variance of the constant reference sample was no less than that for the clinical specimens.

Ovarian cancer pattern diagnostics applied to high-resolution mass spectra

The resulting 216 mass spectra were reanalyzed to generate diagnostic models using the entire range of the aforementioned heuristic parameters. Scatter plots of the total record counts and mean amplitudes of the spectra from the 216 cancer and healthy control samples showed no statistically significant differences in the overall spectra of the two cohorts (Fig. 7) (record count = 359634.2 ± 8223.46, cancer = 354780.9 ± 9813.192; mean amplitude control = 6.018522 ± 1.040222, cancer = 5.204284 ± 1.150888). These values were also statistically

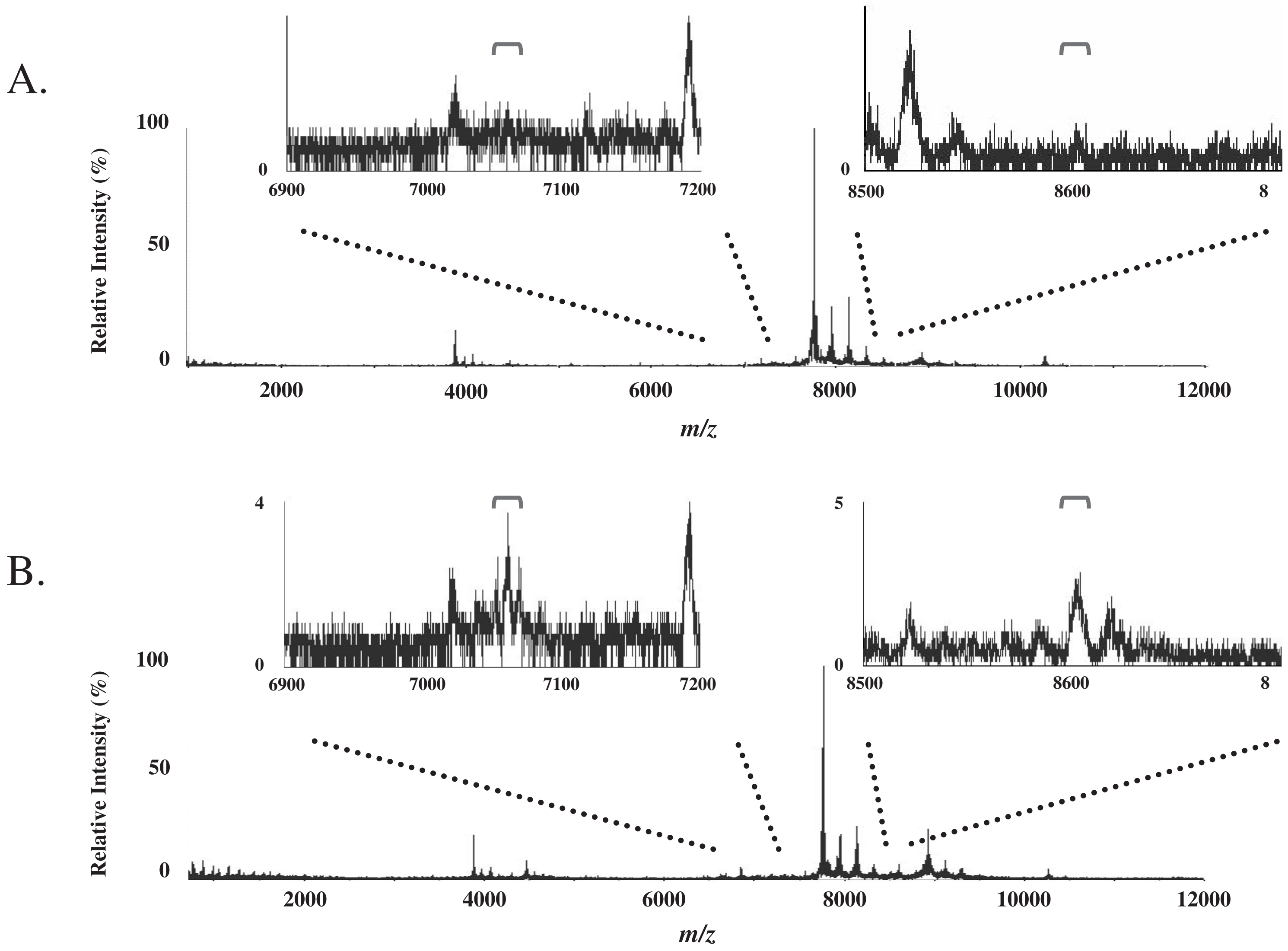


Figure 4 Comparison of SELDI QqTOF mass spectra of serum from an unaffected individual (A) and an ovarian cancer patient (B). Insets show expanded m/z regions highlighting significant intensity differences of the peaks in the m/z bins 7060.121 and 8605.678 (indicated by brackets) identified by the algorithm as belonging to the optimum discriminatory pattern.

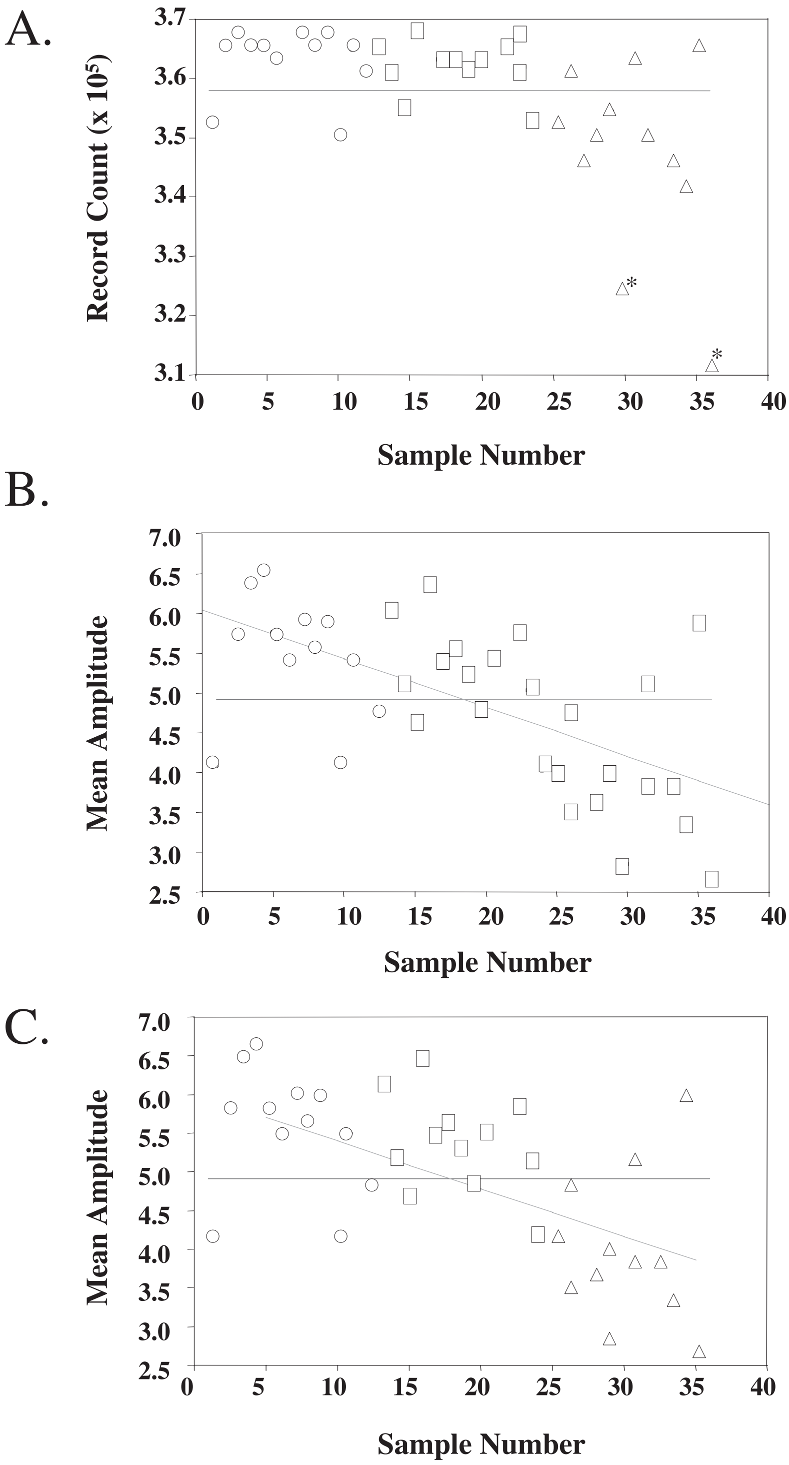


Figure 5 (A–C) Scatter plots of the mass spectra from the serum reference standard co-mingled on the same ProteinChips used for analysis of the 248 serum samples (plotted in Fig. 6). The reference standard was tracked by ProteinChip lot number (B, circles and squares) and day (C). Samples run on day 1 (circles), day 2 (squares) and day 3 (triangles) are indicated.

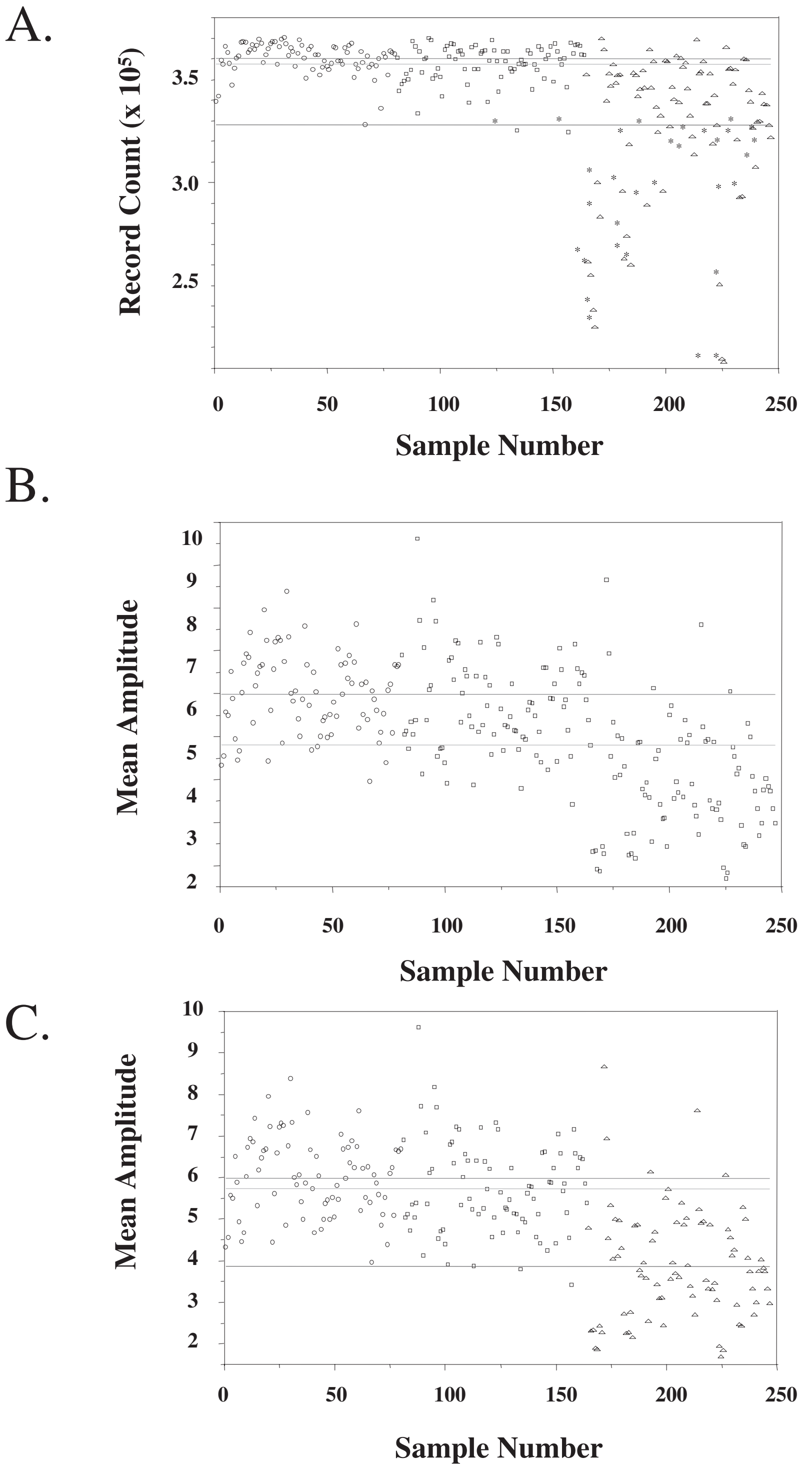


Figure 6 Scatter plots of the mass spectra from the 248 experimental samples representing total mass spectral record count (A) and mean amplitude values. Samples were tracked by ProteinChip lot number (B, circles and squares) and day (C). Samples marked by an asterisk were chosen for exclusion based on the standard deviation (S.D.) of the entire set (greater than 1 S.D. from the mean). Samples run on day 1 (circles), day 2 (squares) and day 3 (triangles) are indicated.

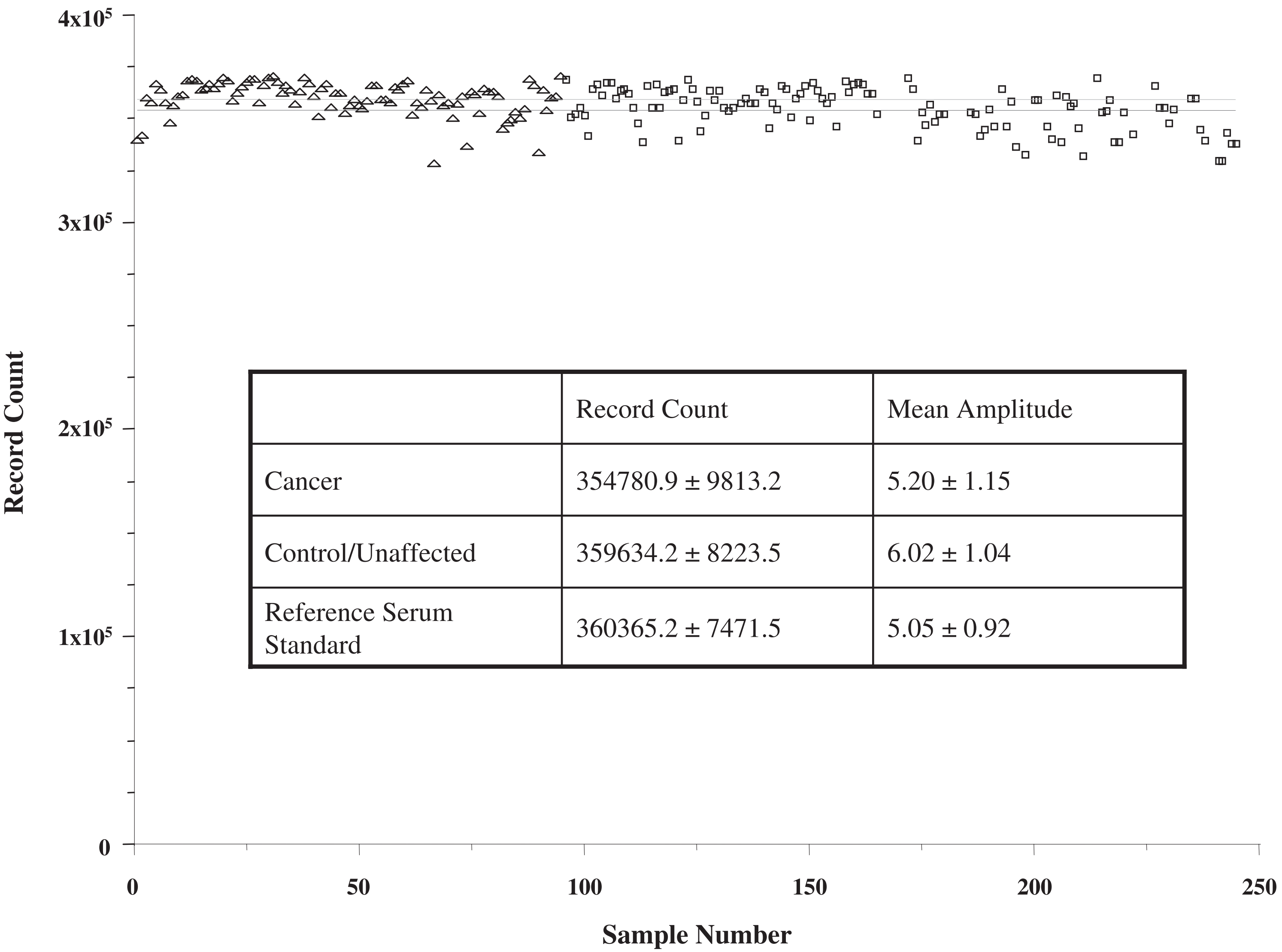
indistinguishable from the reference standard as well (Fig. 7). Models from the training set (52 healthy, 53 ovarian cancer) were validated with a testing set comprising 43 healthy and 68 ovarian cancer serum samples. Not surprisingly, the results of the blinded testing set (Fig. 8) clearly show the ability to discover proteomic patterns in serum analyzed by high-resolution MS. Reflective of the results obtained for the entire sample cohort, four models were found using the reduced data set that were both 100% sensitive and specific, including correctly detecting serum samples taken from all 18 stage I ovarian cancer patients.

Discussion

A limitation of a disease diagnostic that relies on a single biomarker is the lack of sensitivity and specificity when applied to large, heterogeneous populations. Biomarker pattern analysis is an emerging technology aimed at overcoming the limitation of individual biomarkers. While serum proteomic pattern analysis has the potential to provide a new paradigm for diagnosis of early-stage disease, therapeutic monitoring and outcome determination, the success of this method will depend on the ability of a selected set of features to transcend the biologic variability, analytic process variations, and methodologically related background 'noise'. Development of reliable proteomic pattern diagnostic testing for routine clinical use will rely on consistency in sample handling, collection, storage, preservation and shipping procedures. Sample processing and MS instrumentation, software reliability and validation will also greatly influence the ultimate success of this method. One major potential obstacle is the ability of the MS instrument to generate reproducible mass spectra. Drifts in the inherent resolution and mass accuracy within the MS instrument itself are particularly capable of undermining implementation of this new diagnostic in a clinical setting. The more smoothing, warping and filtering required for mass spectral alignment, the greater is the likelihood of reducing the complexity of the data (thus potentially smoothing out real features with diagnostic information) while at the same time generating larger and larger variances between study sets and experiments. It is important to note that the intensities of the selected m/z features are of extremely small amplitude, representing at best only 2% of the intensity of the most abundant peaks present within the serum mass spectra. Some investigators have chosen pattern-recognition methodology which requires a rule-based query at the beginning of the analysis to select, *a priori*, only features that are above a subjective amplitude cutoff (Adam *et al.* 2002, Li *et al.* 2002, Yanagisawa *et al.* 2003). From the present findings, we believe that this approach will delete highly discriminatory features that lie very close to the background noise, as demonstrated in the inset of Fig. 4.

With the exception of the recent study by Tirumalai *et al.* (2003), the low molecular weight (LMW) serum proteome has been relatively unexplored, even though this is the mass region where MS is best suited for analysis. It is possible that disease-associated mass spectra comprise LMW peptide/protein species that may vary in mass by as little as a few daltons. Drifts in resolution and mass accuracy are greatly diminished through the use of a QqTOF MS, allowing for the raw data to be binned. Thus, the same m/z features reproducibly populate the same bins without the need for warping, smoothing, or aligning the raw data.

The mass spectra from the QqTOF MS lead to proteomic patterns with a higher level of diagnostic sensitivity and specificity than those from the lower resolution instrument. Only the QqTOF MS produced mass spectra that resulted in models having 100% sensitivity and specificity, from this particular set of serum samples, for the diagnosis of ovarian cancer. While this diagnostic capability is impressive, it is still critical to the future success of proteomic pattern technology to determine methods to evaluate individual spectral quality prior to, and independent of, bioinformatic classification. These QA/QC tools will not only prevent a potentially erroneous diagnosis based on the use of a 'poor' quality mass spectrum but also allow specific errors in the process to be identified. The QA/QC tools developed in this study serve to (a) evaluate globally sources of variability, (b) identify outliers, and (c) develop specific mass spectral release specifications. By the analysis of reference standards run in parallel to the clinical samples, the major source of mass spectral variability in this study was determined to be instrument related, and not biased by patient phenotype, serum collection or sample preparation methods. Of the 248 initial samples, 216 were 'qualified' by our QA/QC protocol for diagnostic remodeling. Comparisons (t-testing and chi-square analysis) revealed that the variation in the mass spectra (overall amplitude, total record count and deviation in mean amplitudes) between ovarian cancer cases and control samples was statistically indistinguishable from the variance within the process itself, as indicated by the serum reference standard (Fig. 5). Reflective of the analysis of the larger (that is, 248) sample cohort, four models generated using the reduced data set of these 216 statistically 'qualified' samples also attained 100% sensitivity and specificity. While it was not possible for the QA/QC strategy to improve upon the diagnostic capability above that obtained with the original 248 spectra in this



	Record Count	Mean Amplitude
Cancer	354780.9 ± 9813.2	5.20 ± 1.15
Control/Unaffected	359634.2 ± 8223.5	6.02 ± 1.04
Reference Serum Standard	360365.2 ± 7471.5	5.05 ± 0.92

Figure 7 Scatter plots of the serum mass spectra of control individuals (squares) and ovarian cancer patients (triangles) from the 216 spectra after the QA/QC cutoff values were selected. The table inset indicates the mean and standard deviations for the cancer and controls, and the reference standards for both record count and mean amplitude values.

Sensitivity: 100% (68/68 Cancer; 18/18 stage I)
Specificity: 100% (43/43 Healthy)

1	2	3	4
845.089	846.1041	1612.132	845.089
1151.684	998.455	1751.97	982.601
1301.088	1208.302	4413.01	1050.056
2690.843	1936.961	5454.9	1153.989
8516.661	5243.149	7060.121	3974.019
8602.237	7648.014	8602.237	4144.447
8709.548	8622.903	8688.674	7354.07
	8709.548	8709.548	8523.476
	10054.21	9244.304	8602.237
			9510.552

Figure 8 Four distinct models generated from the reduced data set that generated 100% accuracy in testing and validation and the key *m/z* values whose combined intensities discriminated the sets. Common *m/z* features recurring between the models are highlighted.

experiment, it did allow recognition of a process error. Recognition of this error prevented the possible inclusion of poor-quality mass spectra that could potentially affect the bioinformatic discovery of an optimal diagnostic model or result in an incorrect diagnosis.

We hypothesize that diagnostic serum proteomic information, which reflects changes in the physiologic and pathologic state of a target tissue, exist within constellations of small proteins and peptides. We further postulate that serum diagnostic patterns are made up of discrete markers that are a product of the complex tumor–host microenvironment. This hypothesis is supported by the fact that many patterns were found with extremely accurate classification, each of which comprised unique low molecular weight features not found in other patterns as well as low molecular weight species that were repeatedly and consistently found. This result leads to several important implications. Firstly, it is likely that diagnostic proteomic information is partially derived from clipped and/or cleaved host proteins rather than proteins that are directly related to the biology of the tumor itself. Secondly, the biomarker profile may, for example, be amplified by a cascade of systemic processes at the tumor–host interface, resulting in the generation of peptide cleavage products within the tumor–host microenvironment. Thirdly, from a biologic perspective, these assumptions predict the existence of multiple dependent, or independent, sets of proteins/peptides that reflect the systemic response to the regional malignancy. Some recent studies conducted within our laboratory show that cleavage fragments within the LMW range (Tirumalai *et al.* 2003) of the blood proteome contain diagnostic information that is likely to be complexed with larger molecular weight carrier proteins (Mehta *et al.* 2003). This complexation is likely to protect these LMW species from renal clearance, and serve to amplify their overall abundance dramatically. The LMW species and the source of the underlying pattern observed in the present mass spectra may therefore comprise the protein fragments that are bound to carrier proteins.

The data presented here support the existence of multiple highly accurate and distinct proteomic feature sets that can accurately distinguish epithelial ovarian cancer. To screen in a broad population for diseases of relatively low prevalence, such as ovarian cancer, a diagnostic test must exceed 99% sensitivity and specificity, and have clinical utility to reduce the morbidity and mortality associated with this disease. Using and combining multiple diagnostic patterns can possibly overcome the low prevalence to achieve a test with clinical value. In blinded testing and validation, any one of the four best models generated using QqTOF MS data correctly classified serum obtained from epithelial ovarian cancer patients and healthy patients with 100% sensitivity and specificity. Hence, a high-resolution system, such as the QqTOF MS employed in this study, is preferred in view of the present results that serve as a launch-point platform for clinical trials of serum proteomic patterns.

Acknowledgements

By acceptance of this article, the publisher or recipient acknowledges the right of the US government to retain a nonexclusive, royalty-free license and any copyright covering the article. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US government. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract NO1-CO-12400.

References

Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z & Wright GL Jr 2002 Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* **62** 3609–3614.

Agresti A 1990 *Categorical Data Analysis*. New York: Wiley.

Cohen LS, Escobar PF, Scharm C, Glimco B & Fishman DA 2001 Three-dimensional power doppler ultrasound improves the diagnostic accuracy for ovarian cancer prediction. *Gynecologic Oncology* **82** 40–48.

Holland JH 1994 *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA: MIT Press.

Hutchens TW & Yip TT 1993 New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry* **7** 576–580.

Kohonen T 1990 The self-organizing map. *Proceedings of the IEEE* **78** 1464–1480.

Li J, Zhang Z, Rosenzweig J, Wang YY & Chan DW 2002 Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry* **48** 1296–1304.

Mehta A, Ross S, Lowenthal M, Fusaro V, Fishmann D, Petricoin E & Liotta L 2003 Biomarker amplification by serum carrier protein binding. *Disease Markers* **19** 1–10.

Menon U & Jacobs IJ 2000 Recent developments in ovarian cancer screening. *Current Opinion in Obstetrics and Gynecology* **12** 39–42.

Niloff JM, Knapp RC, Schaetzl E, Reynolds C & Bast RC Jr 1984 Ca125 antigen levels in obstetric and gynecologic patients. *Obstetrics and Gynecology* **64** 703–707.

Ozols RF, Rubin SC, Thomas GM & Robboy SJ 2001 Epithelial ovarian cancer. In *Principles and Practice of Gynecological Oncology*, 3rd edn, pp. 165–182. Eds WJ Hoskins, CA Perez & RC Young. New York: Lippincott, Williams and Wilkins.

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC & Liotta LA 2002*a* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359** 572–577.

Petricoin EF 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM & Kohn EC 2002 Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute* **94** 1576–1578.

Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP & Veenstra TD 2003 Characterization of the low molecular weight human serum proteome. *Molecular and Cellular Proteomics* **2** 1096–1103.

Yanagisawa K, Shyr Y, Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S, Moore JH, Caprioli RM & Carbone DP 2003 Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* **362** 433–439.

Zurawski VR Jr, Orjaseter H, Andersen A & Jellum E 1988 Elevated serum CA-125 levels prior to diagnosis of ovarian neoplasia: relevance for early detection of ovarian cancer. *International Journal of Cancer* **42** 677–680.