

Correcting Common Errors in Identifying Cancer-Specific Serum Peptide Signatures[†]

Josep Villanueva,^{‡,§} John Philip,[‡] Carlos A. Chaparro,[‡] Yongbiao Li,^{||} Ricardo Toledo-Crow,^{§,||}
Lin DeNoyer,[^] Martin Fleisher,[⊥] Richard J. Robbins,[#] and Paul Tempst^{*,‡,§}

Protein Center, Molecular Biology Program, Engineering Resource Laboratory, Department of Clinical Laboratories, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York 10021; Spectrum Square Associates, Ithaca, New York 14850

Received February 14, 2005

'Molecular signatures' are the qualitative and quantitative patterns of groups of biomolecules (e.g., mRNA, proteins, peptides, or metabolites) in a cell, tissue, biological fluid, or an entire organism. To apply this concept to biomarker discovery, the measurements should ideally be noninvasive and performed in a single read-out. We have therefore developed a peptidomics platform that couples magnetics-based, automated solid-phase extraction of small peptides with a high-resolution MALDI-TOF mass spectrometric readout (Villanueva, J.; Philip, J.; Entenberg, D.; Chaparro, C. A.; Tanwar, M. K.; Holland, E. C.; Tempst, P. *Anal. Chem.* 2004, 76, 1560–1570). Since hundreds of peptides can be detected in microliter volumes of serum, it allows to search for disease signatures, for instance in the presence of cancer. We have now evaluated, optimized, and standardized a number of clinical and analytical chemistry variables that are major sources of bias; ranging from blood collection and clotting, to serum storage and handling, automated peptide extraction, crystallization, spectral acquisition, and signal processing. In addition, proper alignment of spectra and user-friendly visualization tools are essential for meaningful, certifiable data mining. We introduce a minimal entropy algorithm, 'Entropycal', that simplifies alignment and subsequent statistical analysis and increases the percentage of the highly distinguishing spectral information being retained after feature selection of the datasets. Using the improved analytical platform and tools, and a commercial statistics program, we found that sera from thyroid cancer patients can be distinguished from healthy controls based on an array of 98 discriminant peptides. With adequate technological and computational methods in place, and using rigorously standardized conditions, potential sources of patient related bias (e.g., gender, age, genetics, environmental, dietary, and other factors) may now be addressed.

Keywords: serum • biomarkers • peptides • patterns • bias • mass spectrometry • peak alignment • specimen collection • sample processing

1. Introduction

Serum proteomics and peptidomics are gaining popularity among oncologists in their quest for cancer biomarkers with high diagnostic accuracy. This is based on the premise that sera from cancer patients contain small proteins and peptides, detectable by mass spectrometry (MS), that reflect the presence and the biology of a tumor. Since the first report¹ that serum polypeptide profiling could serve as a diagnostic method for ovarian cancer, many papers have appeared in the literature

using similar approaches.^{2–5} In a number of these publications, discovery of several biomarker patterns are reported, for various cancers, that have diagnostic sensitivities and specificities approaching 100%. If these early results are as robust and reproducible as they seem, then serum proteomics will undoubtedly attain a prominent and lasting position in the future of cancer diagnostics.

Despite initial excitement, skepticism about the methodology and the results is mounting in the scientific community.^{6–10} The simple fact that different research groups have found different discriminatory markers when analyzing similar samples suggests that the serum proteomics methodology may not be reproducible and/or the MS data have been incorrectly interpreted. Several lines of evidence indicate that uncontrolled variables related to both clinical and analytical chemistry may have tainted the published results. Some of the observed "differences" may represent artifacts of blood sample collection and serum preparation, or storage and handling. In addition,

[†] Part of the Biomarkers special issue.

* To whom correspondence should be addressed: Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021. E-mail: p-tempst@mskcc.org.

[‡] Protein Center.

[§] Molecular Biology Program.

^{||} Engineering Resource Laboratory.

[⊥] Department of Clinical Laboratories.

[#] Department of Medicine.

[^] Spectrum Square Associates.

patient related variables such as gender, age, genetic, environmental, dietary, and psychological factors could introduce bias. As many serum proteomic studies have been conducted outside of traditional peptide chemistry/MS laboratories, key variables affecting sample processing, mass analysis and associated signal processing may have been overlooked. Instead, a 'black box' approach has usually been taken whereby serum samples were analyzed using a turnkey, commercial platform and the mass data obtained for large cohorts processed without adequate alignment of the spectra. Such 'blind' sample processing and analysis, with little or no operator control and without interactive data reduction, may not be the best way to uncover novel disease markers.

Serum peptidomic profiling has also stimulated bioinformaticists to develop models to analyze and compare large numbers of data points. Most groups employ 'training' and validation sets of both cancer and control samples, or at the very least cross-validation among a single cohort of cases and controls,^{1–3,11,12} combined with machine learning methods such as support vector machines (SVM),¹³ K-nearest neighbor (K-NN)¹⁴ or decision trees.^{12,15} Earlier work utilized entire spectra ($\geq 20\,000$ data points) as input data¹ but given the paucity of true peptide peaks, the vast majority of the data points represented instrument noise and made analysis unnecessarily complicated. Extracting 'real' peaks, while relatively simple to the expert eye, has now become the focus of considerable informatics effort.^{7,16,17} Several early reports clearly illustrate that peptide profiling done with minimal knowledge of protein chemistry and MS can easily result in artifactual findings.

We have previously reported a technology platform for the simultaneous measurement of large numbers of serum polypeptides.¹⁸ It uses magnetic bead-based, solid-phase extraction of mainly small peptides and a MALDI-TOF MS read-out. The technology is automated on a liquid handling robot in a 96-well microtiter plate format for throughput and reproducibility. The current version utilizes reversed-phase batch processing, although multidimensional chromatographic procedures could be implemented. The system is intrinsically more sensitive than any surface capture on chips as spherical particles have larger combined surface areas than small-diameter spots. When combined with high-resolution MS, hundreds of peptides are detected in a single droplet of serum. Using our system, we have analyzed several thousand serum samples obtained from cancer patients.

From these observations, and from further studies, we have identified additional uncontrolled variables in the clinical and analytical chemistry components and in MS signal processing (prior to statistical analysis) that have misled us in the early days, as well as the authors of some of the recently published papers, into arriving at falsely positive conclusions. Here, we highlight some of these pitfalls, and offer simple recommendations to avoid or correct systematic bias. We also present an improved spectral alignment program and software tools for data visualization and interactive analysis.

2. Materials and Methods

2.1. Materials. Acetonitrile was obtained from Burdick and Jackson (Muskegon, WI), trifluoroacetic acid (TFA) from Pierce (Rockford, IL). Premade *a*-cyano-4-hydroxycinnamic acid (ACCA) matrix solution was purchased from Agilent (Palo Alto, CA). SiMAG-C8/K superparamagnetic, silica-based particles (≤ 1 micron diameter; 80% iron oxide; nonporous), bearing C8 reversed-phase (RP) ligands, were obtained from Chemicell

Table 1. Protocol for Serum Peptide Sample Preparation

Clinical Services:	
1.	Collect venous blood into one BD Vacutainer SST tube, aka tiger-top (Becton Dickinson # 367988). <i>Fill tube to top</i> (8.5 mL).
2.	Following the manufacturer's instructions: <ul style="list-style-type: none"> •Gently invert the tube 5 times to mix clot activator with blood. •Allow blood to clot for <i>1 hour</i> at room temp (RT) in vertical position.
3.	Place the SST tubes on <i>wet ice</i> in vertical position until transport. <ul style="list-style-type: none"> •Put tubes in color-coded bag (<i>PROTEOMICS</i>). •Put in a cooler for transportation to the Clinical Chemistry Lab. •Delivery should be <i>within 2 hrs</i> after clotting.
Clinical Chemistry Lab:	
4.	Spin SST tubes in centrifuge at 1400–2000 RCF for 10 min, at RT.
5.	Transfer the serum (upper phase) to 4-mL cryovials (Fisher Scientific # 0566966). <ul style="list-style-type: none"> •Tubes labeled in advance. •Volume of serum per cryovial should be approximately 1 mL.
6.	<i>Immediately</i> store all samples at -80°C . <i>Avoid freeze–thawing cycles.</i>
7.	Transport samples on dry ice to the Mass Spectrometry Lab.
Mass Spectrometry Lab:	
8.	LIMS and Barcoding: <ul style="list-style-type: none"> •Enter all sample information into LIMS system before samples arrive. •Barcode nine empty 0.25-mL micro-ependorf tubes for <i>each</i> sample. •Barcode incoming cryovials (see #7). <i>Samples must remain frozen on dry ice.</i>
9.	Store all cryovial samples again at -80°C . <i>Avoid freeze–thawing cycles.</i>
10.	Prepare small aliquots for one-time use: <ul style="list-style-type: none"> •Thaw cryovial samples on wet ice. •Aliquot $9 \times 50\ \mu\text{L}$ of each sample in the 9 micro-tubes. •Put in barcoded freezer boxes. •<i>Immediately</i> return all to -80°C storage. <i>Avoid freeze–thawing cycles.</i>
11.	Randomize all samples to be analyzed and generate layouts for the individual runs.
12.	Solid-phase extraction and MALDI plate spotting: <ul style="list-style-type: none"> •Thaw one $50\text{-}\mu\text{L}$ aliquot of each sample on wet ice. •Put in the correct location of the 96-well tube holder. •<i>Immediately</i> start the automated, solid-phase peptide extraction. •Take MALDI plate to MS instrument and start spectra acquisition <i>within 1–2 hrs</i>.

(Berlin, Germany). Other batches of C8 particles (Chemicell) with variously modified properties have also been used for comparative purposes. Human serum control (# S-7023, lot 034K8937) was obtained from Sigma (St Louis, MO). Serum peptide processing was done in 0.2-mL polypropylene tubes (8-tube strips; or 8×12 -tube 'Temp Plate II') from USA Scientific (Ocala, FL).

2.2. Serum Samples. Blood samples from volunteer subjects with no known malignancies and from patients diagnosed with metastatic thyroid carcinoma were collected following a clinical protocol. All collections were approved by the MSKCC Institutional Review and Privacy Board. Blood samples were collected in 8.5-mL, BD Vacutainer SST 'tiger-top' tubes (Becton Dickinson # 367988, Franklin Lakes, NJ), allowed to clot at room temperature for 1 h, and centrifuged at 1400–2000 RCF for 10 min, at RT. Sera (upper phase) were transferred to four 4-mL cryovials (Fisher # 0566966), ~ 1 mL serum in each, and stored frozen at -80°C until further use. The standard procedure is summarized in Table 1. Upon arrival at the mass spectrometry (MS) lab, the cryovials (source vials) were barcoded using the MSKCC Clinical Proteomics LIMS (see below) and a Z4M barcode printer (Zebra Technologies, Vernon Hills, IL). One cryovial of each sample was thawed on ice and used to generate nine smaller aliquots ($50\ \mu\text{L}$ each) in micro-ependorf tubes that were also barcoded and stored at -80°C in barcoded freezer boxes before analysis. Each serum sample had therefore

been frozen and thawed twice before it was subjected to solid-phase peptide extraction and MS (see below).

In selected cases, the standard blood collection and serum preparation protocol was modified for investigative purposes:

2.2.1. Sera Prepared in 'Red-Top' Tubes. Blood samples were collected from healthy volunteers in 8.5-mL, BD Vacutainer, glass 'red-top' tubes (BD # 366430), and sera were prepared with the same protocol used for the blood collected in SST blood collection tubes (Table 1).

2.2.2. Clotting Times. Blood samples were collected in 8.5-mL, BD Vacutainer SST tubes and allowed to clot at room temperature for 5 min. ($t = 0$), 1 and 5 h. The rest of the blood collection protocol was the same as in Table 1.

2.2.3. Freeze–Thaw Cycles. Frozen sera were thawed on wet ice and then immediately refrozen on dry ice and returned to -80°C for 30 min; this procedure was repeated up to 5 times. Aliquots were taken after each freeze–thaw cycle and subjected to automatic solid-phase extraction and MS.

2.3. Automated, Solid-Phase Peptide Extraction. The reversed-phase (RP) batch extraction protocol is automated using a 'Genesis Freedom 100' (Tecan; Research Triangle Park, NC) liquid handling work station¹⁸. Magnetic beads were generally pulled to the side of the tubes (during 'binding', 'washing', and 'elution'). To this end, magnetic strips were embedded in the 96-tube holding plates, between the rows-of-eight. This device was constructed using epoxy-coated, Neodymium–iron–boron (NdFeB) magnets (2 3/4"/1 1/4"/1/8"; L/W/H) from K&D Magnetics (Boca Raton, FL). To resuspend the beads in a minimal volume of elution solvent, they must first be collected at the bottom of the tubes, requiring the magnets to be positioned directly underneath in a 96-well micro titer plate layout. A plate holder containing 96 NdFeB-magnetic disks, 1/4" in diameter and 1/4" thick (obtained from Force field; Fort Collins, CO), was therefore constructed. A 96-well plate cooler (Eppendorf, Westbury, NY) was incorporated into the system to reduce solvent evaporation.

The Gemini 100 system was programmed either directly via its standard software or, when individual wells needed to be accessed independently, indirectly through its work-lister capability. Once in the 96-well format, this system automates all of the liquid-handling steps, including magnetic separation via a robotic manipulating arm, mixing of eluates with MALDI matrix and deposition onto the Bruker 384-spot MALDI target plates.

2.4. MSK Clinical Proteomics LIMS. The large number of samples that can now be processed by the automated system creates problems in locating and accessing both the physical samples and their associated data. In collaboration with the North Shore Long Island Jewish Health System (Manhasset, NY) biorepository group, we have developed a complete Laboratory Information Management System (LIMS). By characterizing and standardizing the entire sample collection and serum preparation process, we designed a SQL server database with a webpage based interface. This LIMS provides (i) a physical sample inventory control system with bar coding capabilities for easy identification of samples, (ii) a database storing clinical and nonclinical sample information, and (iii) the generation of 'AutoX' files, necessary to operate the AutoFlex MALDI-TOF mass spectrometer automatically through the 'AutoXecute' function. The web interface enables platform independent entry and access to data and other control features. Together, these capabilities make the LIMS a complete information

solution. It tracks the handling of samples from storage to the final mass spectra.

2.5. Mass Spectrometry. Peptide profiles were analyzed with an Autoflex MALDI-TOF mass spectrometer (Bruker; Bremen, Germany) equipped with a 337 nm nitrogen laser, a gridless ion source, delayed-extraction (DE) electronics, a high-resolution timed ion selector (TIS), and a 2 GHz digitizer. The standard protocol, except where modified for investigative and optimization purposes, was as follows. Separate spectra were obtained for two restricted mass-to-charge (m/z) ranges, corresponding to polypeptides with molecular mass of 0.7–4 kDa (" $\leq 4\text{kD}$ ") and 4–15 kDa (" $\geq 4\text{kD}$ ") (assuming $z = 1$), under specifically optimized instrument settings. Each spectrum was the result of 400 laser shots, per m/z segment per sample, delivered in four sets of 100 shots (at 50-Hz frequency) to each of four different locations on the surface of the matrix spot. The effective laser energy delivered to the target was carefully controlled to be 16 μJ ($\pm 10\%$) per shot (see below). A weekly performance test with a control commercial human serum is done and the effective laser energy delivered to the target is modified accordingly. The entire irradiation program was automated using the instrument's 'AutoXecute' function. Spectra were acquired in linear mode geometry under 20 kV (18.6 kV during DE) of ion accelerating and -1.3 kV multiplier potentials, and with gating of mass ions ≤ 400 m/z ($\leq 4\text{kD}$ segment) or ≤ 3000 m/z ($\geq 4\text{kD}$ segment). DE was maintained for 80 ($\leq 4\text{kD}$) or 50 nanoseconds ($\geq 4\text{kD}$) to give appropriate time-lag focusing after each laser shot. Peptide samples were always mixed with two volumes of premade α -cyano-4-hydroxycinnamic acid (ACCA) matrix solution (Agilent; Palo Alto, CA), deposited onto the stainless steel target surface, in every other column of the 384-spot layout, and allowed to dry at room temperature. Thirty fmols (per peptide) and 500 fmols (per protein) of commercially available calibration standards (Bruker Daltonics # 206195 ($< 4\text{kD}$) and # 206355 ($> 4\text{kD}$)) were also mixed with ACCA matrix and separately deposited onto the target plates, adjacent to each spotted serum sample (one sample/one standard), in the alternating columns.

2.6. Laser Pulse Energy Measurements. The Bruker AutoFlex MALDI-TOF instrument does not provide a direct measurement of the laser energy that is delivered to the sample. Instead, it features an uncalibrated setting ('control bar') of the laser power, that ranges from 0 to 100% on an arbitrary scale without providing an absolute value. The instrument does have a power/energy monitoring probe at the output of the nitrogen laser, and the beam is attenuated in the path to the sample by means of a matched pair of counter-rotating optical flats. The attenuation control provided to the user rotates the two flats from a position of maximum reflectance (set to 0%) to a position of maximum transmittance (set to 100%).

The nitrogen laser (MNL 200-C; Lasertechnik, Berlin, Germany) in the AutoFlex instrument has a manufacturer specified, nominal output energy of 105 μJ per pulse. The value reported by the laser monitor of the instrument, before the attenuator, was 108 μJ . We measured the laser energy per pulse, by integrating 100 pulses, with a J9-484 pyroelectric joulemeter probe (18.8 VJ^{-1} responsivity) (Molelectron Detector Inc., Portland, OR) before and after the attenuator set at maximum transmittance (100% on the control bar). Our readings were 1.8 mV and 1.47 mV, respectively, indicating that the effective maximum transmittance of the attenuator is only 82%. The difference between our measurement with the pyroelectric detector (1.8 mV/18.8 $\text{VJ}^{-1} = 96 \mu\text{J}$) and that of the laser monitor

(108 μ J) can be attributed to a lack of accurate calibration of our detector. Nonetheless, it is the relative measurements (before/after the attenuator) that are relevant.

We then measured the laser energy after the attenuating device as a function of the user controllable attenuator settings. These measurements were done with a USB2000 fiber coupled spectrometer (Ocean Optics, Dunedin, FL). This instrument was used because it provides good sensitivity and signal-to-noise to measure the 337.1 nm laser line down to the 20% attenuator setting, as well as providing a convenient optical fiber probe to position inside the instrument. The measurements were normalized and scaled to the maximum transmittance value of 82%. The laser setting that had been empirically determined was now measured in this way to yield 16- μ J energy per pulse, post-attenuation. Since we now know all the interconversion factors, we can readily calculate the laser energy delivered to the target per shot by a simply reading the instrument's laser monitor. This is done once a week and adjustments are made accordingly to compensate for fading laser energy over time.

2.7. Signal Processing. 2.7.1. Setup. Once acquired, the data are stored with a naming convention that allows each sample to be associated with its calibrant. The spectra are converted from Bruker's binary format to ASCII files containing two columns of data (x : m/z , y : intensity) by a custom written macro in FlexAnalysis (Bruker). For the lower mass range (700–4000 Da), about 48 000 x,y points were generated while for the upper mass range (4–15 kDa), there were about 77 000 points.

Further data processing was done in MATLAB (see below) with a custom script called 'Qcealign' using only the ASCII versions of the raw spectra. 'Qcealign' used the 'Qpeaks' program (Spectrum Square Associates, Ithaca, NY) for smoothing, baseline subtraction and peak labeling. For these samples, the singlet width parameter required by 'Qpeaks' was set to –400 for the lower mass range and –200 for the upper mass range, thereby specifying the resolution, $(m/z)/\Delta(m/z)$, for processing. This peak information was used automatically by 'Qpeaks' in setting the parameters for smoothing, baseline-subtraction, and binning. The noise statistics were set to 'Normal'.

Since each parameter has tremendous importance in the processing, we created custom software written in MATLAB to load and view individual spectra and examine the impact of changing individual parameters. This software, called Signal Processing & Preview (SPP), is a graphical viewer of the spectral curves from ASCII files. It functions as an interface for DSP (Digital Signal Processing) subroutines (provided by 'Qpeaks'), while hiding the technical details from the users. It can plot raw and processed spectra side-by-side to review the outcome of signal processing. The parameters of the DSP module can be adjusted, if necessary, and the results are updated immediately.

2.7.2. Processing. Once the parameters are selected, a setup file is created. 'Qcealign' queries the setup file to obtain a list of all the directories to be processed. During a single processing run, all data files in all listed directories are aligned with each other. For each directory, singletwidth information is provided in the setup file, along with parameters controlling calibration, peak labeling sensitivity, alignment, etc. In 'Qcealign', the files containing the polypeptide standards are calibrated first. The centroid positions of peaks in these calibration files are obtained from the peak table created by 'Qpeaks' compared to the known polypeptide peak positions, and a quadratic calibration equation for correcting the measured masses in

each calibration file is created. The calibration equations are saved to disk for use in calibrating the mass axes of the sample files.

Next, 'Qcealign' creates a reference file to which all sample spectra will later be aligned. The first data file is loaded and calibrated by applying the curve calculated from its associated calibrant spectrum. This file's x -axis (m/z) becomes the x -axis (and thus the calibration) used in the reference file. 'Qcealign' then loads all other sample files, calibrates them, and adds their intensities to the reference file's intensity. Calibrating the other samples is important since it brings the samples' peaks closer to the position of the peaks in the calibrated reference file. Once all samples are added, the reference spectrum becomes the average of all the sample files. Peaks in the reference spectrum are slightly broadened, compared to peaks in individual spectra, because external calibration is imperfect. Alignment, an internal calibration step, would refine the external calibration.

The reference is processed with 'Qpeaks' to find a baseline which is then subtracted. The reference is normalized to unit size by dividing each intensity value by the Total Ion Count (TIC). Once normalized, a scaling factor is added by multiplying each intensity value by a user-selected number (e.g., 10^7). This scaling factor is constant within a data set and is used primarily to set the normalized spectrum to a more "user-friendly" scale, where most peak heights are greater than one. Next, 'Qcealign' processes each sample file with 'Qpeaks' to create a peak table, smoothed curve and a baseline. This spectrum is then passed on to alignment.

2.7.3. Alignment. 'Qcealign' aligns each spectrum to the reference file, using a custom alignment algorithm called "Entropycal". The alignment algorithm slides the data file by ' n ' data points to the right or left along the x -axis of the reference file. At each relative position n , the Shannon entropy¹⁹ of the sum of the two files is computed. The alignment position is the shift n that produced the minimum Shannon entropy of the configuration. The data file is then returned with the x -axis shifted. These data shift is equivalent to adjusting the zero-point of time in the TOF spectrometer. Once aligned, the smoothed spectrum, which is produced as a byproduct of the 'Qpeaks' processing, is then updated to reflect the aligned m/z values and saved to disk after baseline-subtraction, normalization and scaling. This final processed ASCII data is later visually inspected in a custom built Mass Spectra Viewer.

The previously generated peak table is updated to reflect the baseline-subtraction, normalization, scaling and alignment. The peak information is added to a column of a master matrix. Rows in the matrix represent integer m/z values; peaks are written to the row closest to the calibrated m/z position of the peak; columns represent samples. This matrix is needed later for the 'binning' process. Separate peaklists with the peak intensities and m/z values are saved to disk. This process from the invoking of 'Qpeaks' to the saving of peak lists is repeated for all samples. After all sample files are processed, the master matrix of peak positions is queried, and positions of peaks in the matrix are 'binned' to create a spreadsheet that can be used for statistical analysis. This binning is done using the declared resolution $(m/z)/\Delta(m/z)$; all peaks in rows within $\Delta(m/z)$ of the strongest peak at a given value of m/z are binned together. Finally, the binned results are written to disk.

To study the impact of this methodology, we analyzed the same data in two additional ways. In the first, we processed raw spectra without calibration and alignment. In the second, we added calibration, leaving out the alignment step. The

results of all three analyses were then taken for similar statistical analysis to examine the impact of each step. To implement the two additional analyses, a custom MATLAB script was written called “process_and_save.” This script is an earlier variant of ‘Qcealign’ that does not include alignment or binning. Using this script, the first additional analysis was processed without calibration. ‘Qpeaks’ was used with the same parameters as described earlier. Baseline-subtraction, smoothing, normalization and scaling was also done. The results of these two additional steps were aligned by using a custom-made File-maker Pro (Clarix Corporation, Santa Clara, CA) database to transform the peaklists into a spreadsheet format for statistical analysis. The peaklists were imported into the database and peaks from all related samples were then binned at 1500 ppm, as described.¹⁸ Peaklists from the aligned samples were also binned using 1500 ppm, allowing us to compare binning using ppm and singletwidth.

2.8. MATLAB Software Tools. MATLAB (MathWorks, Natick, MA) was chosen as a developing platform because of its high-level programming language for quick deployment of applications. The current version contains three modules of visualization and analysis tools:

2.8.1. Signal Processing & Preview. SPP functions as an interface for signal processing subroutines (Qcealign), while hiding the technical details from the users. SPP works in either Single Mode or Batch Mode. In the single mode, raw and processed spectral curves will be displayed side by side for review. The investigator loads one or more representative spectral data, fine-tunes the parameters of the DSP subroutine, and examines the outcomes. Whenever processed data is not available, the program will automatically call ‘Qcealign’ to generate them. Furthermore, the user can modify the DSP parameters that are passed on to ‘Qcealign’. After settling on optimal parameters, a script will be generated automatically to work in the batch mode, which processes large number of spectral data in one command.

2.8.2. Mass Spectra Viewer. MSV is a visualization interface for processed spectral data. MSV plots the spectra as *X*–*Y* curves (mass vs magnitude) for examining the signatures of several groups of samples. It provides flexible browsing functions such as scroll, zoom, highlighting, etc. Besides visualization, the second critical design is to provide the data management for querying and categorization of large numbers of samples. We incorporated the database design, which uses a unique identification number (Sample_ID) for each entry. All attributes associated with the spectra are identified by this ID. Users can then search and classify single samples or groups of samples based on their properties. The third aim of this program is to handle large data sets (thousands of spectra) with good responsiveness during loading, visualizing and saving procedures. This is achieved by merging all processed spectra, along with the unique Sample IDs, into an internal multidimensional data storage (matrix). Disk operations are very rapid because, instead of loading/writing thousands of spectra, MSV reads from or writes to a single file. The speed of browse function is also improved by plotting all the spectral curves once and then rescaling the axis of the window. In this way, stress on the video card and memory bus is reduced, because only a small portion of data needs to be rendered. The optimization is further enhanced by optional Resolution Control, which allows the user to control the number of points per pixel plotted in the graph. For Matlab programs, these optimizations improve performance significantly.

2.8.3. HeatMap. HM is a module specifically designed to display the binned spreadsheet of spectral peaks coming from the DSP subroutines. Peaks are plotted as a 2D heat map image with sample names and normalized ion intensity values as the *X*- and *Y*-position coordinates, in which the magnitudes of the *m/z* are color coded on a gradient scale. In addition to the ordinary browsing functions such as zoom and scroll, the user has the option to reorganize the rank of *X*- and *Y*-position coordinates, without the constraints of statistical correlation that are enforced by most heat map commercial software packages.

2.9. Statistical Analysis. The binned spreadsheet, containing data from spectra obtained for all samples of cancer patients or healthy subjects (59 samples total; 549 *m/z* values, with normalized intensities for each sample; >33 000 data points), was imported into the ‘GeneSpring’ program (Agilent; Palo Alto, CA) and analyzed using various statistical algorithms such as ANOVA, PCA, hierarchical clustering, K-NN and SVMs. In ‘GeneSpring’, an “experiment” was created to represent the masses. No normalizations were applied to the experiment since the masses were normalized by Qcealign. In the parameter section of the experiment, a parameter called “Cancertype” was created to label samples as either Thyroid cancer or Control. In Experiment’s Interpretation section, the Analysis mode was set to “Ratio (signal/control)” and all measurements were used. No Cross-Gene Error model was used either.

2.9.1. Mann–Whitney U Test. Once the experiment was created, the *m/z* values (“peaks”) were filtered by using a nonparametric test (Mann–Whitney U test). The Benjamini and Hochberg method was used to adjust *p*-values for multiple comparisons. The threshold for significance was an expected false discovery rate less than 1×10^{-5} . These tests are meant to find peaks that show statistically significant differences between the two groups studied.

2.9.2. Hierarchical Clustering. All the masses as well as the ones selected from the Mann–Whitney test were then subjected to clustering analysis using the available tool in ‘GeneSpring’. The peaks were organized by creating mock-phylogenetic trees (dendrograms) called “gene trees” and “experiment tree” in the software. The trees were displayed with the samples along the *X*-axis and the masses along the *Y*-axis. The clustering method for both trees also measured similarity by Standard Correlation (also known as “Pearson correlation around zero”).

2.9.3. Class Prediction. K-nearest-neighbor (K-NN) analysis was done by using the Class Prediction Tool in ‘GeneSpring’. The training group was set to experiment created earlier. The Parameter to Predict was set to Cancertype. The Gene selection was set to use all 549 masses. The number of neighbors was set to five with a *p*-value decision cutoff of 1. The K-NN was used to cross-validate (using the ‘leave one out’ cross-validation method) the training set. The SVM was done with the same training set and parameters and set to cross-validate. The kernel used was polynomial dot product (Order 1) with a diagonal scaling of 0.

3. Results and Discussion

3.1. Serum Preparation and Handling. A comparison of different serum preparations revealed that the type of tube used for blood collection, the clotting times and clotting temperatures all have a significant impact on the mass spectra of peptide populations extracted from several cohorts of samples. Our results showed that “biomarker patterns” with diagnostic

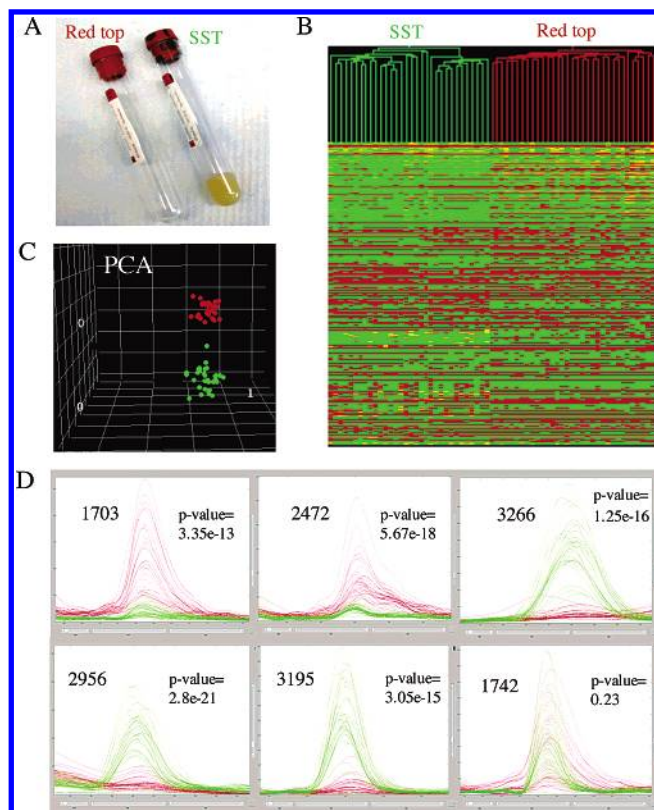


Figure 1. Effect of the blood collection tube on MS-based serum peptide profiling. Blood from 32 healthy volunteers was collected in both red-top and SST tubes following the protocol described in Table 1. A. Red-top and SST tubes. B. Hierarchical clustering of 32 serum samples prepared in both red-top (red) and SST tubes (green). All 608 m/z peaks were used in the comparison. C. Principal Component Analysis (PCA) of the 2×32 samples. D. Overlay of mass spectra obtained from two groups of 32 samples. All spectra were processed as described in the methods section, and are displayed using Mass Spectra Viewer (MSV) (see Methods). The centroid of the bin and its Mann–Whitney adjusted p -value to separate the two groups (red-top vs SST) are shown for each peak.

accuracies of 100% in leave-one-out cross validation (LOOCV; using k-NN or SVM) analysis could be discovered when comparing the spectra generated from two identical sets of 32 different blood (control) samples prepared in either BD glass ('red-top') or BD SST ('tiger-top') tubes (Figure 1A). As further shown in Figure 1 (panels B and C), unsupervised cluster analysis and principal component analysis (see 'Methods' and the 'GeneSpring' user manual) allowed easy segregation of the two sample sets. A number of m/z values ('peptides') exhibited remarkable "biomarker" potential; five of those (all with adjusted p -values $< 1 \times 10^{-13}$) are shown in Figure 1D as color-coded spectral overlays, together with one example ($m/z = 1742$; p -value = 0.23) of other peptide peaks that did not segregate the groups.

In a separate study, four blood samples from different persons were collected in SST tubes and were allowed to clot at RT for 5 min ($t = 0$), 1 h and 5 h, followed by sample processing, peptide extraction and MS. As illustrated in Figure 2 for five different m/z peaks, intensities did vary in a time-dependent manner, likely reflecting changes in abundance. Sometimes intensity diminished ($m/z = 1062, 1077, 1352$), but in other cases it increased ($m/z = 1703, 2272$). A possible

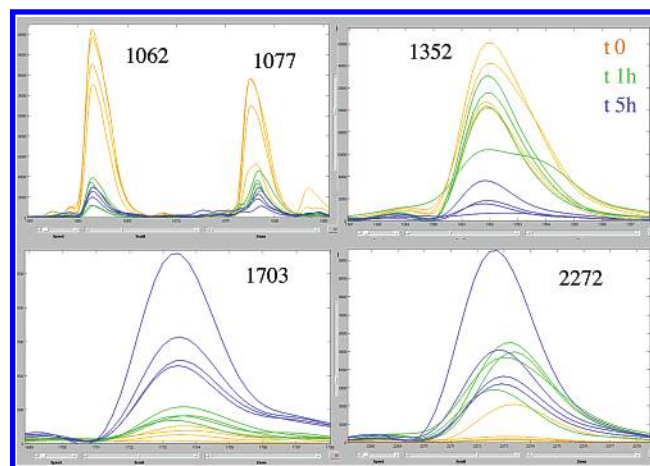


Figure 2. Effect of clotting time on MS-based serum peptide profiling. Blood samples from 4 healthy volunteers were collected in 8.5-mL, BD Vacutainer SST tubes and allowed to clot at room temperature for 5 min (orange), 1 h (green) and 5 h (blue). The rest of the blood collection protocol was the same as in Table 1. All spectra were processed as described in the methods section, and are displayed using Mass Spectra Viewer (MSV) (see Methods). The centroid of the bin is shown for 5 representative m/z peaks.

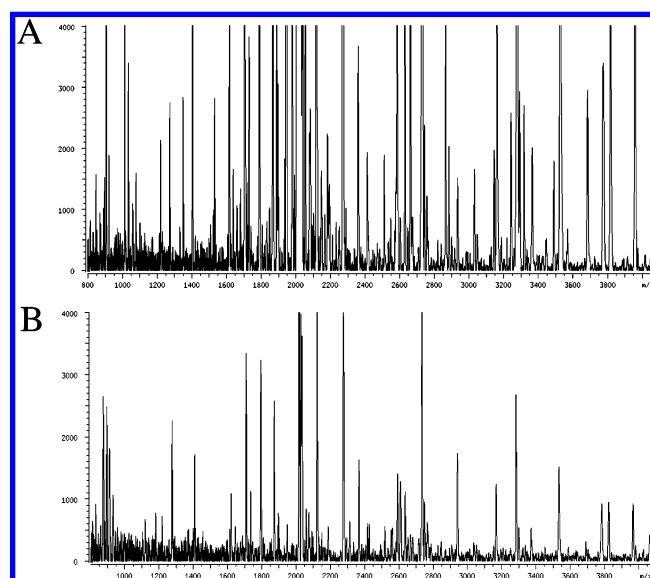


Figure 3. Effect of freeze–thaw cycles on serum peptide profiling using RP magnetic particles and MALDI TOF MS. Frozen sera (at -80°C) were thawed on wet ice, sampled and immediately placed at -80°C for 30 min; this was repeated 4 times. Aliquots were taken after each freeze–thaw cycle and analyzed by the automated RP-protocol and MALDI-TOF MS. A. Mass spectrum from serum submitted to 2 freeze–thaw cycles. B. Mass spectrum from serum submitted to 4 freeze–thaw cycles.

explanation is the degradation of plasma peptides or, on the other hand, formation/accumulation of new peptides during and after the clotting process.

Storage details of the samples are also critical. As shown in Figure 3, the number of freeze–thaw cycles of sera (e.g., 2 versus 4) dramatically affects the resulting patterns observed by MS, most likely due to peptide aggregation, precipitation and adsorption to surfaces. At our institution, serum samples are always frozen and thawed twice; the second thawing step

immediately before peptide extraction and MS analysis (see Table 1).

In sum, any systematic bias in serum preparation and/or storage between two or more groups of samples can result in a statistically relevant, yet clinically useless diagnostic tool. All the above examples illustrate the importance of strict adherence to a standardized protocol when test samples from patients and healthy controls have been obtained from different sources. We have therefore made a concerted effort to instruct nurses, phlebotomists, messenger service staff, and clinical technicians about the importance of strictly adhering to the standard protocol.

3.1.1. Automated Solid-Phase Extraction. We also tested how peptide extraction might affect the final results. Capturing serum peptides on either a (hydrophobic) SELDI chip or RP-magnetic particles is a critical process. We have previously shown the effects of carbon chain length and the sample-to-magnetic bead ratio (v/w).¹⁸ We have since observed that in order to obtain total reproducibility, extraction of the peptides must be done using the exact same batch of magnetic particles and in a totally automated, randomized fashion. The spectra shown in Figure 4 are derived from six identical aliquots of a single serum sample, processed separately using the standard automated protocol but with different batches of Chemicell SiMAG-C8/K particles (arbitrarily designated C8_1 to 6). Note that the tests were done under collaborative agreement with the manufacturer and that none of the particles used here are any longer commercially available. However, our laboratory has acquired bulk quantities of the 'C8_1' batch, and those beads have been used throughout all the tests and analyses described herein. As illustrated in Figure 4, profound differences were observed in peak patterns. We highly recommended that anyone planning to use solid-phase RP extraction of serum peptides for profiling purposes verify potential batch-dependent differences of the particles of choice. This also applies to SELDI-TOF MS; any one project should be done in its entirety with the same lot of chips and the serum samples should ideally be processed automatically as well.

After previously incorporating a liquid handler for fully automated analyses, serum samples from patients and control subjects were typically processed in consecutive runs and/or were dispensed in the 96- or 384-format microvials, and similarly applied to the target plates, arrayed as 'blocks'. We noticed that this approach could also lead to systematic bias (data not shown). In our current procedure, a randomization program (part of the LIMS) is used to position case and control samples for processing. There are also no noticeable day-to-day differences, as illustrated by the experiment shown in Figure 5. Seven samples, taken from a single batch of commercially available human serum, were analyzed in the presence of calibration standards over the course of seven weeks; each sample every week, all randomized on the plate. Samples analyzed on any particular day were then treated as a group during subsequent data processing and nonsupervised hierarchical clustering. Color-coding in Figure 5A,B is by day and, clearly, there is no bias. All colors are scrambled in the dendrogram (Figure 5A) and in the single, tight cluster resulting from PCA (Figure 5B). Spectral overlay analysis of selected peaks further suggests the absence of day-to-day bias (Figure 5C).

3.1.2. MALDI-TOF Mass Spectrometry. Another factor affecting the results is the actual acquisition of the mass spectra. Once the serum peptides are captured, they are eluted and

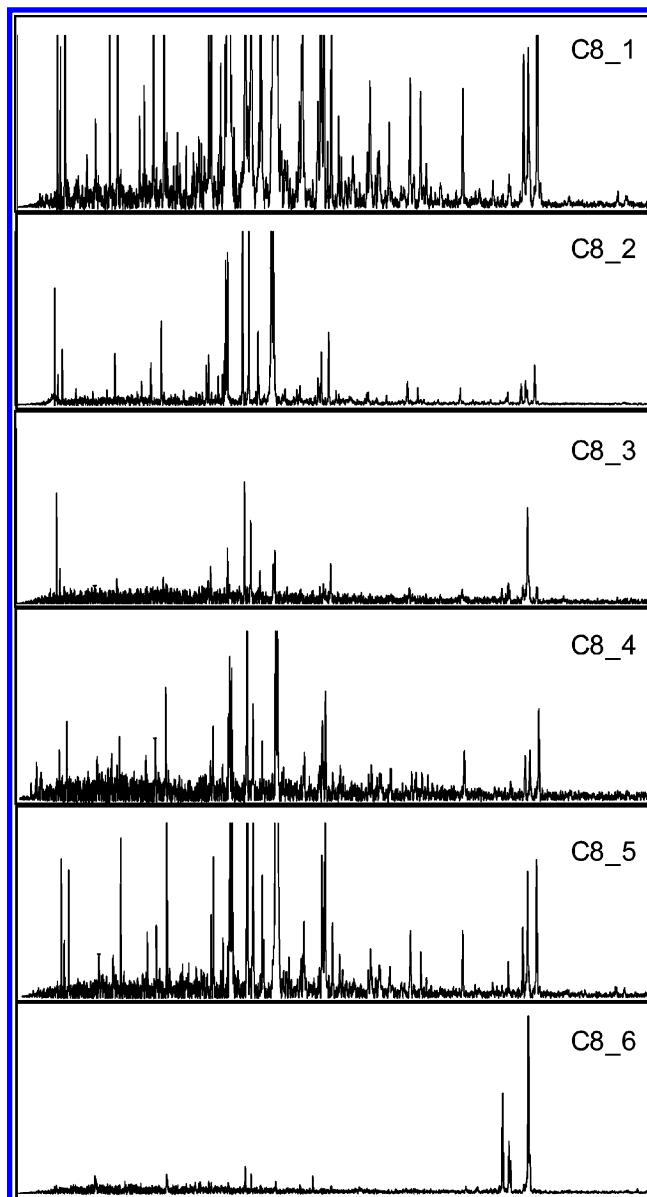


Figure 4. Effect of different batches of C8/K magnetic particles on serum peptide profiling. Equal volumes of serum were incubated with fixed-weight amounts of different batches of SiMAG-C8/K RP beads (with variously modified properties) in separate experiments. Beads were washed, eluted with 50% acetonitrile in water, and the eluates analyzed, all as described under 'Methods'. Segments of the MALDI-TOF mass spectra corresponding to peptides in the 0.7–4 kDa mass range (assuming $z = 1$) are shown.

mixed with a UV-absorbing matrix in a fashion conducive to laser-induced ionization and desorption. To this end, cocrystals must be formed that are suitable for the production of detectable peptide ions upon irradiation with a pulsed nitrogen laser. This process is not entirely understood and must therefore be optimized empirically for each type of sample.

Matrix–solvent composition has a major effect on sample crystallization for MALDI, organic solvent type/concentration and pH being the most important features²⁰. Different solvent compositions have been previously evaluated, including varying concentrations (in water) of MeCN, 2-propanol and mixtures of MeCN/methanol, all with and without 0.1% TFA. We found

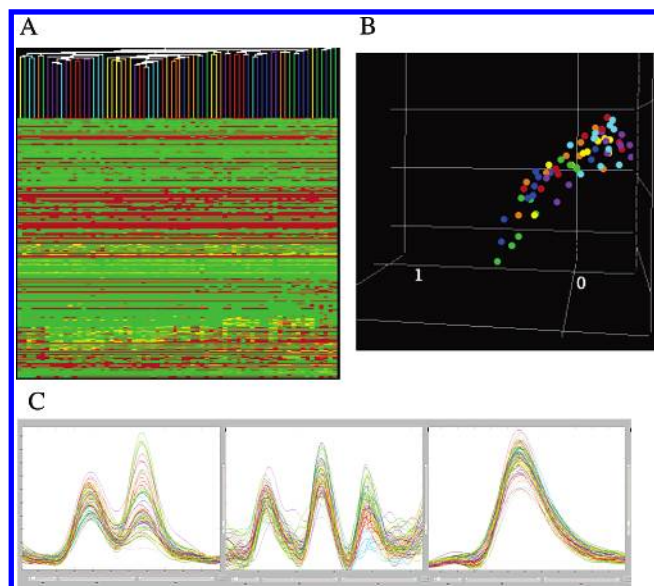


Figure 5. Reproducibility of automated, solid-phase peptide extraction and MALDI-TOF MS. Seven runs consisting of 10 aliquots of the same serum sample randomized over a 96-position micro-tube holder were done using the TECAN liquid handler. Samples were independently processed and analyzed over seven consecutive weeks, using SiMAG-C8/K magnetic beads and the standard analytical protocol (see 'Materials and methods'). A. Hierarchical clustering was done on the seven runs, whereby each run is represented with a different color. All 553 m/z peaks were used in the comparison. B. Principal Component Analysis (PCA) of the seven runs using all the m/z . Colors are the same used for hierarchical clustering. C. Overlay of mass spectra obtained from the seven runs done over seven consecutive weeks. All spectra were processed using the signal processing described in the methods section, and are displayed using Mass Spectra Viewer (MSV) (see Methods).

that mixtures of 40% acetonitrile/50% methanol/10% water, without TFA, gave the maximum number of serum peptide peaks in the spectra. This composition is similar to a commercially available, premade matrix solution (Agilent) which, in our hands, gives consistently satisfactory spectra. Another key variable is crystal 'stability'. We have noticed an adverse effect of increasing the time between crystallization and mass spectra acquisition. While we have no ready explanation for this unexpected observation, we can easily control the effect by taking all spectra within a maximum of 1–2 h after completion of robotic sample processing.

The number of laser shots averaged for each mass spectrum, which determines the signal-to-noise ratios for the serum peptide ions, must also be pre-selected. From examining 36 published studies all utilizing SELDI-TOF to develop diagnostic patterns of disease, it appears that the number of shots averaged to produce a spectrum varied from 50 to 192, delivered in 1 to 13 sets. Consequently, although these studies all used the same biological fluid, they are virtually incomparable. In addition, the authors never mention whether the samples from the different case and control groups had been randomized in space (e.g., on the target plates/strips; see above) and over time for unbiased mass analysis. Each of the above could potentially have skewed the data.

We conducted several experiments to evaluate the effects of the number of averaged laser shots and the laser energy delivered to the target per shot. Figure 6 gives the summary of

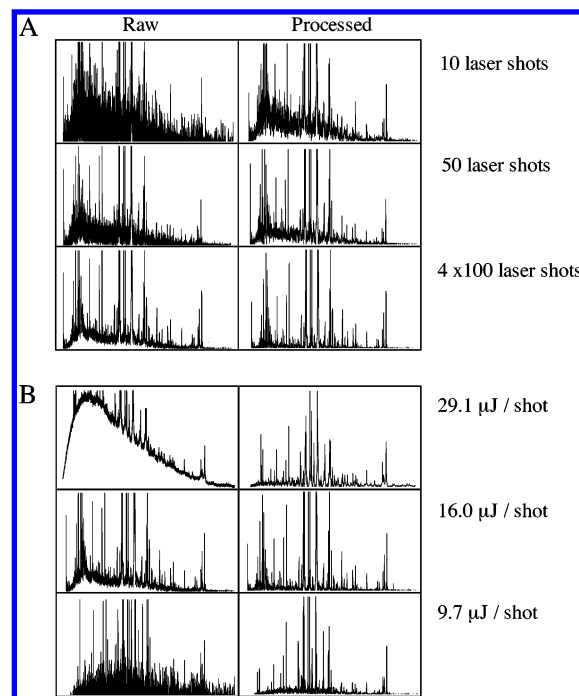


Figure 6. Effects of MALDI-TOF mass spectral acquisition variables on serum peptide profiling. Equal volumes of serum were incubated with fixed-weight amounts of magnetic SiMAG-C8/K beads, beads washed, eluted with 50% acetonitrile in water, and prepared for MALDI-TOF MS. Spectra were acquired in linear mode geometry. Segments of the MALDI-TOF mass spectra corresponding to peptides in the 0.7–4 kDa mass range (assuming $z = 1$) are shown. All the spectra derive from the same sample, at different locations in a single deposit on the target. All the spectra were processed using the signal processing described in the methods section. A. Three spectra were generated averaging different number of laser shots (indicated). The raw and processed spectra are displayed using Mass Spectra Viewer (MSV) (see Methods). B. Three spectra were generated using different effective laser energy (indicated) delivered to the target. Both the raw and processed spectra are displayed using Mass Spectra Viewer (MSV) (see Methods).

an extensive investigation, by comparing the empirically determined optimal set of conditions (4 × 100 laser shots; 100 each in 4 different locations on the sample spot; laser shots delivered on the target at 16 μJ ($\pm 10\%$) per shot—see below) with suboptimal ones. Insufficient numbers of laser shots resulted in very noisy spectra, even after smoothing and baseline subtraction (Figure 6; top panel; 'processed'). More than 100 laser shots per location depletes the crystal and is therefore ineffective; hence the 4 × 100 arrangement. Averaging still more laser shots, e.g., 1000 (10 × 100) or more, did not further improve the quality of the spectra (data not shown). Note that the spectra in Figure 6 have been scaled to the same size; in actuality, signals in the spectra resulting from averaging 10 or 50 laser shots were much lower than those obtained under standard conditions. Somewhat different results were obtained when changing the laser energy, where both higher and lower settings than the empirically determined optimal conditions resulted in deterioration of the spectra (Figure 6; bottom panel). This observation has some consequences for the way laser energy should be controlled. The Bruker AutoFlex MALDI-TOF instrument (like most others) does not have a direct read-out of the laser energy delivered to the target; instead it has probe

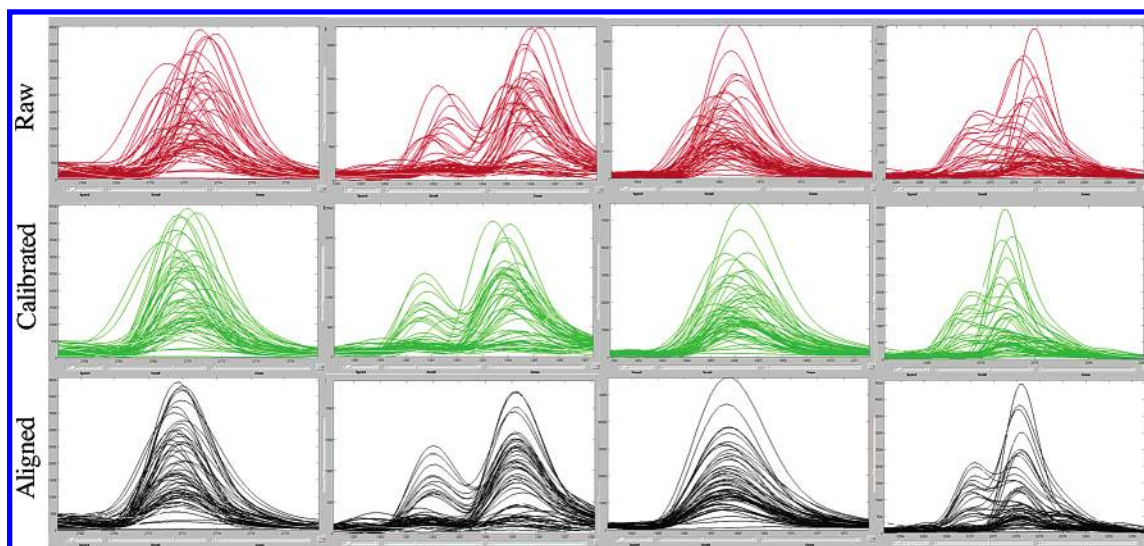


Figure 7. Effects of mass calibration and 'Entropycal'-based alignment on mass spectra overlays. Fifty nine mass spectra obtained from serum samples of control individuals and thyroid carcinoma patients are shown in overlay. All spectra have been smoothed and baseline subtracted using the signal processing described in the methods section. Different mass calibrations were applied. Four regions of the spectra were selected and they are displayed using Mass Spectra Viewer (MSV) (see Methods). Each mass spectrum region is shown in a raw version (Raw), after external calibration (Calibrated) and after external calibration plus computer 'Entropycal' alignment (Aligned). External calibration and 'Entropycal'-based alignment are described in the methods section.

at the output of the laser, before the attenuator. We verified the accuracy of this monitoring device and then calibrated the settings of the attenuator (displayed on the computer screen as an arbitrary scale of 100–0%) by measuring transmitted energy at varying %. This allowed us to generate a calibration curve to convert before-to-after attenuation laser energy. Laser output energy is measured and documented on a weekly basis. The gradual decrease in energy over time is then offset by selecting a higher setting on the control. For example, the setting had to be increased from an arbitrary 30% to 40% over a six month time period to maintain an effective laser energy of $\sim 16 \mu\text{J}$ at the target plate.

Molecular mass calibration is another key issue. In nearly all published reports on MS-based biomarker discovery, the authors used some version of external calibration but exact details are not mentioned. In one particular report, it is stated that the external calibration was done on a weekly basis. In our experience, it is simply not possible to have 0.1% (1000 ppm) mass accuracy when analyzing human serum peptides in the linear mode of MALDI-TOF, as stated in most of the SELDI-TOF papers, unless external calibration is done for *each* sample by depositing peptide standards on the nearest-neighbor spot location (see 'Methods'). The effect on subsequent, uncorrected alignment in spectral overlays is clearly illustrated in Figure 7 (top two panels), and it further improves the results of computational alignment as well (see below).

3.1.3. Signal Processing. The final step before statistical analysis is signal processing of the mass spectra. There are several components to this step: smoothing, baseline correction, normalization, calibration/alignment, and peak labeling. As long as the same signal processing is applied to all the samples in a data set, there will be no systematic bias to confound the statistical analysis. However, it is possible that suboptimal signal processing will suppress or obscure the true signal/information in the data.

One means to optimal signal processing is to employ software in which an expert user (with a priori knowledge of

what the data should look like) can fine-tune as many parameters as possible. This may lead to good results but could be time-consuming and would be difficult to reproduce without expertise. Another approach is to use an algorithm that requires few inputs from the user and makes few assumptions about the data. The software that we used here, 'Qpeaks' (see also under 'Methods'), finds peaks in a spectrum and outputs various information, including a peak table, smoothed trace and an optional baseline. This table contains peak parameters (apex positions, centroids, widths, areas), and standard errors in the parameters. The peak-finding function *rzrpic* is based on a Bayesian second derivative of the data. Smoothing is accomplished with the Maximum Entropy (Maxent) smoothing function *rzresm*. Both the Bayesian second derivative and the Maxent smoothing are optimal in that they produce the best possible results using only a single assumption: namely, a width for peaks in the data ('singletwidth'). They do not require any additional parameters, such as polynomial degree used by Savitsky-Golay algorithms. Bayesian and Maxent algorithms are self-adjusting to the noise, so no "degree of smoothing" parameter is required. This handling of noise means that the Bayesian and Maxent functions cannot compromise resolution by oversmoothing. Thus, the use of primarily one parameter simplifies data processing without compromising data quality. Additionally Qpeaks works well with low resolution data such as the kind acquired in MALDI TOF linear mode, unlike many other standard peak labeling algorithms which are optimized for isotopically resolved spectra.

Once processed, the spectra need to be aligned to compare similar peptides across samples, an operation that may be the most difficult task in peptide profiling studies. A common approach is to perform external calibration; i.e., peptides of known molecular mass are analyzed alongside the samples. A calibration curve is then calculated to adjust the *x*-axis of the calibrant spectrum so that its known peaks fit their known values. This curve can be applied to spectra of samples that were spotted in nearby positions on the MALDI target. Ideally,

each sample position on the target plate has an adjacent calibrant spot. But despite the best possible external calibration, m/z peaks representing identical peptides in different samples deviate to various extents from the theoretical molecular mass. They are slightly shifted to the left or right in the spectra, which makes strict 'numerical' alignment (e.g., a spectrum divided in 1-Da, consecutive segments) all but impossible. Usually to correct this shift, a mass percentage window (estimated accuracy of the data set, e.g., 0.15%) is applied along the mass range instead, and common peaks across samples are binned in this way. As a result, a composite peak list comprising all samples is generated for statistical analysis. Detailed investigation revealed, however, that even such spectral alignments can be the source of artifacts or noise, in particular leading to an inflated number of peaks.

We found that a better approach to alignment is to apply a function, termed 'Entropycal', which aligns sample data files to a reference file using a minimum entropy algorithm, and by taking unsmoothed ('raw'), baseline-corrected data. Using raw spectra for the alignment allows all the statistical information in the data to be used. Processed data contains less information. The alignment is performed in three steps: reference spectrum creation, 'Entropycal' and binning. First, a reference spectrum is created by summing all the intensities of all the calibrated samples together. This results in a composite spectrum that contains the average of the peak information from all the data sets. The x -axis of the reference spectrum is the x -axis of the first calibrated sample. Next, 'Entropycal' slides each data file by n data points to the right or left along the x -axis of the reference file. At each relative position n , the Shannon entropy of the sum of the two files is computed. The optimal alignment occurs at the shift that produces the minimum Shannon entropy. Third, the aligned peak lists are then binned by using the resolution of the peaks: all peaks in rows within $\Delta(m/z)$ of the strongest peak at a given value of m/z are binned together, and a spreadsheet is created for further statistical analysis. Further details can be found in the 'Methods' section.

This approach appears to complement the signal processing of 'Qpeaks'. Again, no a priori information is assumed and no expert fine-tuning is required. In fact, no user intervention is required as both reference spectrum and alignment parameters are calculated on the fly. With an improved alignment (see Figure 7; bottom panel) and binning, fewer (but higher quality) peaks are passed on to statistical analysis. Further investigation revealed that with this signal processing and alignment, less noise and fewer artifacts are passed on to the statistical analysis. We will illustrate the benefits of these signal processing routines with an example, discussed below, that involved analysis of two small cohorts (cancer and control) of serum samples.

3.1.4. Serum Peptide Profiling to Detect Thyroid Carcinoma. Most thyroid cancer survivors are in surveillance programs to detect tumor recurrences. The MSKCC Endocrine Service performs in-depth evaluations on more than 350 high-risk patients each year involving physical exams, blood work, ultrasounds, and a variety of radiological procedures. Over the past 10 years, on average, metastatic disease is found in approximately one-third. Blood samples are routinely drawn for a tumor marker (thyroglobulin). In a preliminary study, we collected sera from 27 patients who had clear evidence for residual metastatic (MET) thyroid carcinoma and from 32 healthy volunteers (mixed gender; ages between 22 and 50). Specimens are linked to database records but were anonymized

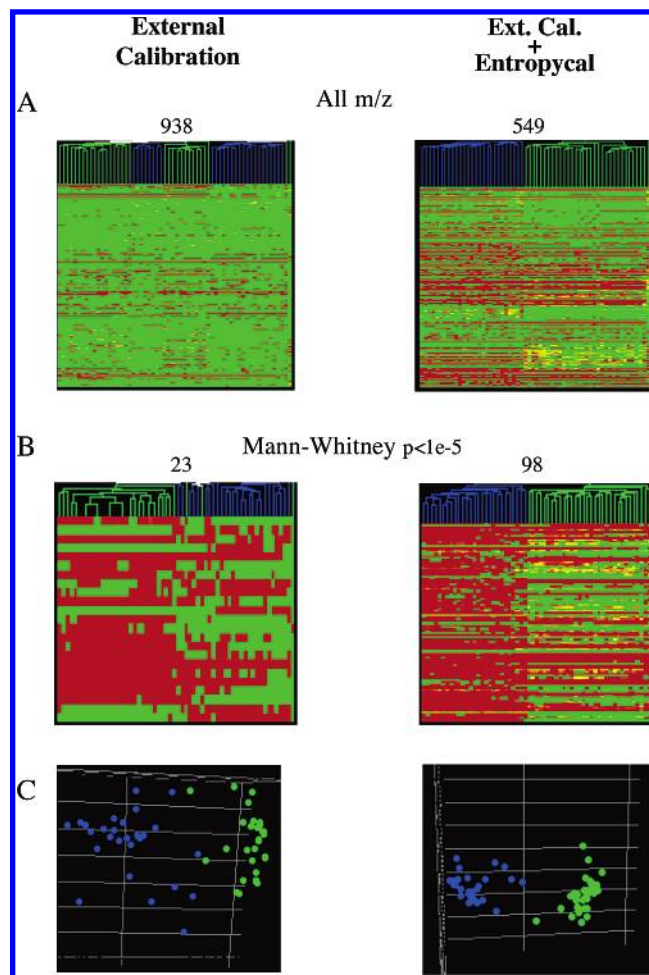


Figure 8. Effect of mass calibration on statistical analysis. Spreadsheets containing binned peaklists for 59 mass spectra obtained from sera samples of 32 control individuals (green) and 27 thyroid carcinoma patients (blue) were imported into the Genespring program. Two spreadsheets contained a binned peaklist after smoothing, baseline subtraction and normalization described in the methods section. However, mass calibration and binning were different in the two sets: data with external calibration and binning at 1500 ppm (Calibrated) and data with external calibration and alignment using 'entropycal' (Aligned). A. Hierarchical clustering using all the bins was done on the two different signal-processing sets. B. Hierarchical clustering done using the m/z that passed the Mann-Whitney set at $p < 1 \times 10^{-5}$ for the two sets: Calibrated and Aligned. C. Principal Component Analysis (PCA) of the two sets using only the m/z that passed the Mann-Whitney test at $p < 1 \times 10^{-5}$.

and stripped of any patient identifiers to meet HIPPA guidelines. We have analyzed the peptide patterns in these sera to determine if they can identify those who harbor metastatic lesions with high diagnostic accuracy.

Blood collection, sera preparation, storage and handling, automated peptide solid-phase extraction and mass spectrometry were exactly as listed in Table 1, as described in the various 'Methods' and 'Results' sections herein, and as shown in the diagram in Figure 10 depicting the entire serum peptidomics operation. The main objective was to explore the impact of signal processing on statistical analysis. First, using the 'standard' approach, external calibration and peak binning across all 59 spectra, using an m/z tolerance of 0.15%, resulted in 938 unique bins or 'peaks'. The resulting spreadsheet contains

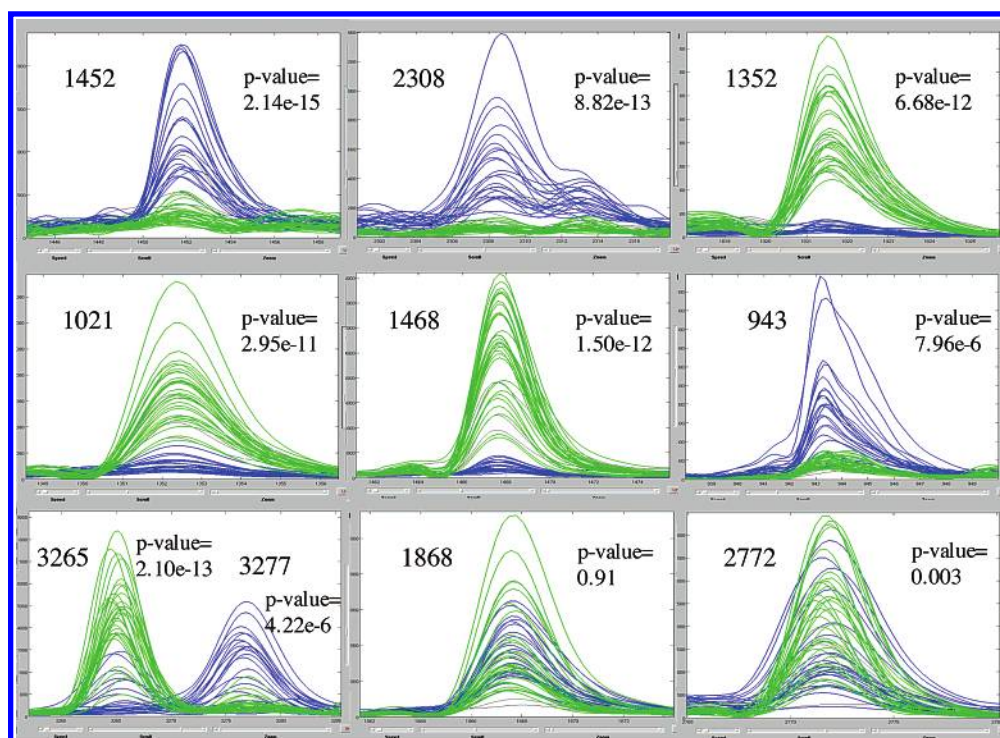


Figure 9. Serum peptide signatures of thyroid tumor patients. Overlay of mass spectra obtained from serum samples of 32 control individuals (green) and 27 thyroid carcinoma patients (blue). All spectra were processed using the signal processing described in the methods section. Some m/z with either very small or large p -values, calculated using the Mann–Whitney test, were selected and are displayed using Mass Spectra Viewer (MSV). The centroid of the peak and its Mann–Whitney p -value to separate the two groups (controls and thyroid carcinoma) are shown for each peak.

55 342 entries listed as normalized intensities that vary over roughly 4 orders of magnitude on an arbitrary scale. The readout is similar to that of microarray gene expression profiling and commercial ‘GeneSpring’ software was therefore used to evaluate peptidomic data. The spreadsheet was used to create a hierarchical cluster, using the Pearson correlation (Figure 8A; left side panel), and for principal component analysis (PCA; data not shown). A fairly good, but not perfect, segregation of ‘Thyroid Cancer’ and control samples was obtained. We then performed a nonparametric Mann–Whitney test at the $p < 1.0 \times 10^{-5}$ level. Interestingly, 23 unique peptides ($\sim 2.5\%$ of total) passed this selection whereas less than one was expected by chance only ($938 \times 1.0 \times 10^{-5}$). To get a visual representation of the signal strength, these results were also taken for hierarchical clustering and PCA analysis (Figure 8B and C; left side).

When calibrated spectra were additionally subjected to the ‘Entropycal’ alignment and binning function, all peaks across all 59 spectra could be merged into just 549 m/z bins. The resulting data set allowed near perfect segregation of the Thyroid cancer and control samples without any prior statistical test to select the most discriminatory peaks; i.e., in a totally unsupervised manner (Figure 8A; right side panel). Furthermore, 98 peptides ($\sim 18\%$ of total) passed a Mann–Whitney test at $p < 1 \times 10^{-5}$. Clustering analysis and PCA using this highly selected set of peaks (Figure 8B,C; right side) resulted in a markedly improved result as compared to a data set obtained by classical binning using $\% (m/z)$ tolerances. We conclude that a much higher percentage (18 vs 2.5%) of highly distinguishing spectral information is retained after a Mann–Whitney test of data sets derived from well-aligned peak patterns. By contrast, some of these peaks become spread out

over two to three bins during ‘standard’ binning, thereby diluting marker potential. This was confirmed by comparing the hierarchical clustering results for the two approaches (Figure 8A).

To individually verify the above findings, the 98 selected m/z peaks were visually inspected in color-coded overlays (using the MSV) of all spectra of both sample cohorts. As can be readily observed for 8 out of the 10 selected examples shown in Figure 9, peak intensities may differ dramatically between the two groups; sometimes consistently higher in all the cancer samples, sometimes higher in all controls. The corresponding adjusted p -values are all below 1.0×10^{-5} . Two peaks shown in Figure 9, at $m/z = 1868$ and 2772 , were not part of the selected group of 98 and have little to no biomarker potential. Note that some of the peaks shown in Figure 9, are also shown in Figure 7, both in aligned and unaligned (‘raw’) overlays.

4. Conclusions

‘Molecular signatures’, a recent concept in biology and medicine, are derived from qualitative and quantitative patterns of entire groups of biomolecules (e.g., mRNA, proteins, peptides, or metabolites) in a cell, tissue, biological fluid or an organism. To apply this concept to biomarker discovery, the measurements should ideally be noninvasive and performed in a single read-out. In a previous report, we described a prototype, MS-based technology platform for automated, solid-phase extraction of peptides from biological fluids, coupled to a direct MALDI-TOF read-out.¹⁸ We have since explored whether this platform can reproducibly define peptide patterns in human serum to look for ‘signatures’ that may indicate

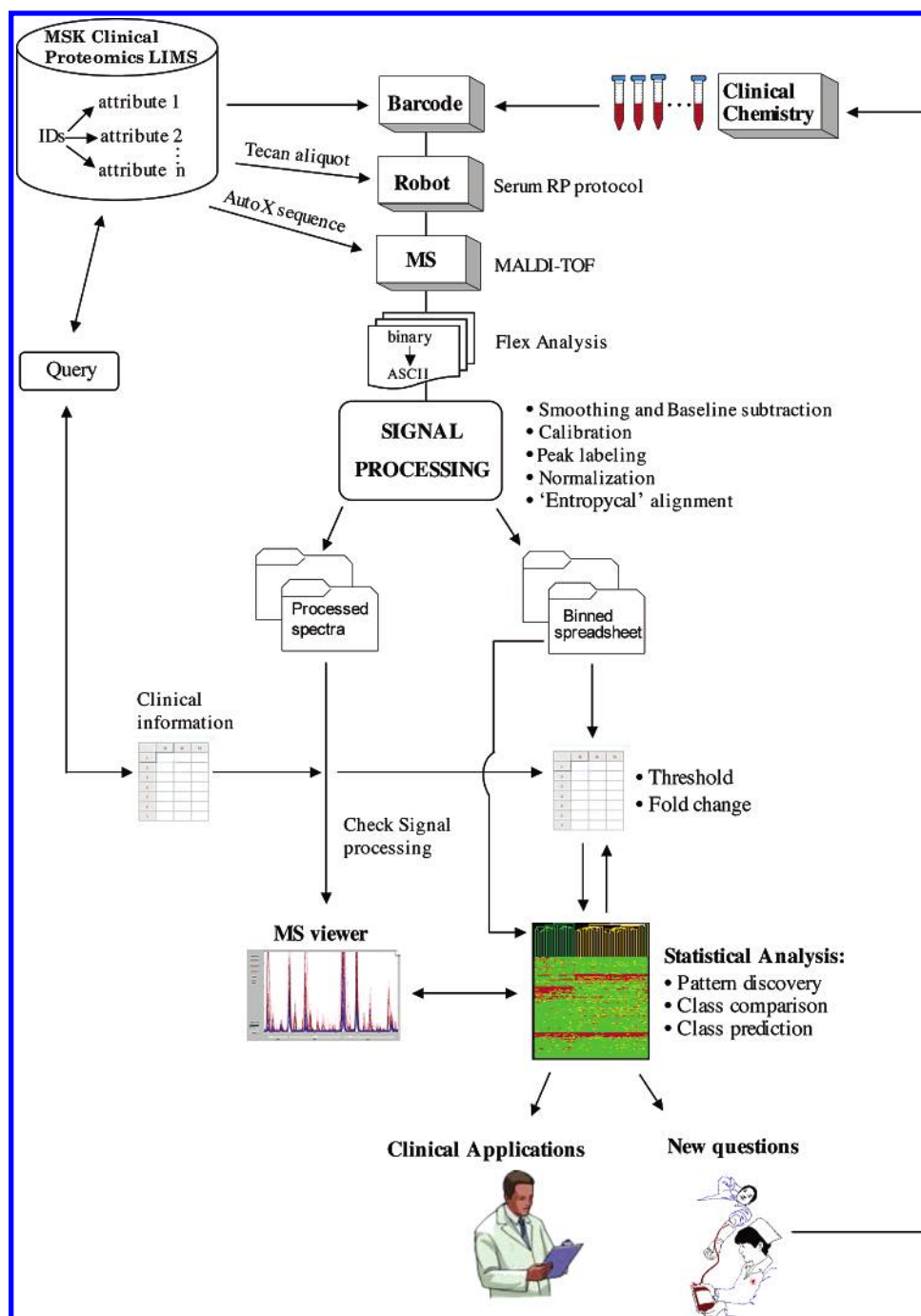


Figure 10. MSKCC serum peptidomics operation. Flowchart showing the various steps of the serum peptidomics approach described in this report.

disease, such as the presence of cancer. While many analytical variables had already been optimized and initial tests suggested possible diagnostic potential, recent observations by us and others have discovered previously unrecognized sources of bias.^{6–10}

Diagnostic peptide profiling must be rigorous, extremely reproducible and free of bias to have any future clinical utility. We have now evaluated, optimized, and standardized a number of variables that may affect final spectral patterns; from specimen collection, storage and handling, throughout all steps of analytical chemistry and MS signal processing. Also, proper alignment of spectra is essential to have any chance of subsequent meaningful data mining. In fact, it makes biom-

arker pattern detection a task for which common statistical tools are as good or even better suited than some of the recent pattern recognition algorithms. Overall, the results have guided us to outline a uniform sample preparation protocol (Table 1), to develop adequate infrastructure, procedures, and software tools (Figure 10), and to establish expertise, training, and organizational structure within Memorial Sloan-Kettering's serum peptidomics program.

Using this improved analytical platform and a commercial statistics program, we found that sera from patients with metastatic thyroid carcinoma can be distinguished from healthy controls based on an array of 98 discriminant peptides. This limited application provides further proof that sera from

patients with solid tumors contain peptides, detectable by MALDI-TOF MS, that reflect presence of cancer. We speculate that serum peptide signatures similarly derived from much larger study sets could eventually rival diagnostic accuracy of conventional cancer biomarkers. As we will increase the stringency of statistical tests to select peptides that discriminate between groups, fewer peptides will meet the criteria, but it should push diagnostic accuracy higher. For instance, to distinguish cancer subgroups or stages, to define the rate of disease progression, or to predict and monitor responsiveness to therapy, including recurrence.

To completely validate the prospect of serum peptide profiling as a means to diagnostic marker or marker pattern discovery, another major source of potential bias must first be analyzed. Many patient related variables, such as gender, age, genetic factors, drug treatment, and environmental, dietary, behavioral, and psychological factors have never been systematically investigated. It is also not known whether patterns are unique to each individual, and whether they can change with physiological events or stay constant over time. We believe that with the adequate technological and computational means as well as the expertise now in place, and when using rigorously standardized conditions, these issues can be addressed in relatively short order. The ultimate proof of the value of this new approach will be in the ability of other laboratories to reproduce either the entire peptidomic patterns or to show that the highly discriminatory peptides have the same amino acid sequences.

Acknowledgment. We are indebted to Andrew Martorella for assistance, and to Eric Holland, David Shaffer and Hans Lilja for the gift of control specimens. We also thank them and other colleagues in the Protein Center and the Clinical Services for support and discussion. Supported by the Himanshu Vakil Research Fund, an ABC2 award and Developmental Funds from NCI Core Grant P30 CA08748.

References

- (1) Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A. *Lancet* **2002**, 359 (9306), 572–577.
- (2) Adam, B. L.; Qu, Y.; Davis, J. W.; Ward, M. D.; Clements, M. A.; Cazares, L. H.; Semmes, O. J.; Schellhammer, P. F.; Yasui, Y.; Feng, Z.; Wright, G. L., Jr. *Cancer Res.* **2002**, 62 (13), 3609–3614.
- (3) Li, J.; Zhang, Z.; Rosenzweig, J.; Wang, Y. Y.; Chan, D. W. *Clin. Chem.* **2002**, 48 (8), 1296–1304.
- (4) Borozdenkova, S.; Westbrook, J. A.; Patel, V.; Wait, R.; Bolad, I.; Burke, M. M.; Bell, A. D.; Banner, N. R.; Dunn, M. J.; Rose, M. L. *J. Proteome Res.* **2004**, 3 (2), 282–288.
- (5) Ebert, M. P.; Meuer, J.; Wiemer, J. C.; Schulz, H. U.; Reymond, M. A.; Traugott, U.; Malfertheiner, P.; Rocken, C. *J. Proteome Res.* **2004**, 3 (6), 1261–1266.
- (6) Diamandis, E. P. *Mol. Cell Proteomics* **2004**, 3 (4), 367–378.
- (7) Baggerly, K. A.; Morris, J. S.; Coombes, K. R. *Bioinformatics* **2004**, 20 (5), 777–785.
- (8) Check, E. *Nature* **2004**, 429 (6991), 496–497.
- (9) Ransohoff, D. F. *Nat. Rev. Cancer* **2005**, 5 (2), 142–149.
- (10) Villanueva, J.; Tempst, P. *Nature* **2004**, 430 (7000), 611.
- (11) Rosenblatt, K. P.; Bryant-Greenwood, P.; Killian, J. K.; Mehta, A.; Geho, D.; Espina, V.; Petricoin, E. F.; Liotta, L. A. *Annu. Rev. Med.* **2004**, 55, 97–112.
- (12) Qu, Y.; Adam, B. L.; Yasui, Y.; Ward, M. D.; Cazares, L. H.; Schellhammer, P. F.; Feng, Z.; Semmes, O. J.; Wright, G. L., Jr. *Clin. Chem.* **2002**, 48 (10), 1835–1843.
- (13) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. *Bioinformatics* **2000**, 16 (10), 906–914.
- (14) Wagner, M.; Naik, D.; Pothén, A. *Proteomics* **2003**, 3 (9), 1692–1698.
- (15) Wu, B.; Abbott, T.; Fishman, D.; McMurray, W.; Mor, G.; Stone, K.; Ward, D.; Williams, K.; Zhao, H. *Bioinformatics* **2003**, 19 (13), 1636–1643.
- (16) Zhu, W.; Wang, X.; Ma, Y.; Rao, M.; Glimm, J.; Kovach, J. S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, 100 (25), 14666–14671.
- (17) Yasui, Y.; Pepe, M.; Thompson, M. L.; Adam, B. L.; Wright, G. L., Jr.; Qu, Y.; Potter, J. D.; Winget, M.; Thornquist, M.; Feng, Z. *Biostatistics* **2003**, 4 (3), 449–463.
- (18) Villanueva, J.; Philip, J.; Entenberg, D.; Chaparro, C. A.; Tanwar, M. K.; Holland, E. C.; Tempst, P. *Anal. Chem.* **2004**, 76 (6), 1560–1570.
- (19) Shannon, C. E. *Bell System Tech. J.* **1948**, 27, 379–423.
- (20) Cohen, S. L.; Chait, B. T. *Anal. Chem.* **1996**, 68 (1), 31–37.

PR050034B