Serum Peptidome Patterns That Distinguish Metastatic Thyroid Carcinoma from Cancer-free Controls Are Unbiased by Gender and Age*§

Josep Villanueva‡§, Andrew J. Martorella¶, Kevin Lawlor‡, John Philip‡, Martin Fleisher∥, Richard J. Robbins¶**‡‡, and Paul Tempst‡‡§ §§

Serum peptidomics is a special form of functional proteomics. The small number of blood proteins that are the source of most prominent peptides in human serum serve as a substrate pool for commonly occurring and/or cancer-derived proteases. Exoprotease activities in particular, when superimposed on the ex vivo coagulation and complement degradation pathways, contribute to generation of not only cancer-specific but also “cancer type”-specific serum peptides. Following development of a unique, semiautomated serum peptide profiling platform and after completing investigations to eliminate common experimental bias, we have now studied possible effects of gender and age on serum peptidomes of 200 healthy men and women, ages 20–80, and of 60 patients (30 men and 30 women) with metastatic thyroid carcinomas. Extensive MALDI-TOF MS and data analysis suggested negligible contributions of both age and gender to the serum peptidome patterns except that healthy men and women under 35 years, but not older individuals, could be distinguished with ~70% accuracy. Considering the more advanced age of most patients, this finding is unlikely to interfere with peptidomics analysis of most cancers. By examining patient samples and age/gender-matched controls followed by variability analysis of either demographic or disease (versus control) groups, we could conclusively rule out demographic bias. An optimized, 12-plexion thyroid cancer signature was then developed, enabling classification of an independent validation set with 95% sensitivity and 95% specificity (binomial confidence intervals, 75.1–99.9%). Ten of these peptides had previously been assigned to signature patterns of other solid tumor cancers. One of the two newly discovered peptides was dehydro-Ala³-fibrinopeptide A. As we expand this study to include hundreds of thyroid cancer patients, the peptide signature will be adjusted, further validated, and then evaluated in a clinical setting used either independently or in combination with existing markers. Molecular & Cellular Proteomics 5: 1840–1852, 2006.

Serum peptidomics has generated considerable excitement among oncologists and analytical chemists in recent years (1–5) as this technology may hold promise to establish rapid cancer blood tests through mass spectrometry-based profiling of patient proteomes/peptidomes. However, some concerns have also been voiced regarding biological, technological, and data mining artifacts that may introduce bias (6). The most common sources of bias in serum proteomics-based biomarker discovery are related to sample collection, processing and storage, patient demographics (e.g. gender and age), analytical chemistry, and data analysis methods. Follow- ing the introduction of a new platform to semiautomatically and simultaneously measure serum peptides by utilizing magnetic, reversed-phase beads for analyte capture and a MALDI-TOF MS readout, several sources of bias related to clinical and analytical chemistries, mass spectrometry, sample handling, and spectral analysis have already been studied and remedied (7, 8). For instance, blood collection tubes, clotting times and temperature, and the number of freeze-thaw cycles are all critically important as well as surface chemistries and batch-to-batch variation of the magnetic particles, MALDI sample crystallization, laser irradiation, and automation of the solid-phase extraction, which were all major sources of variation (7, 8). In addition, spectral alignment appeared to be the most challenging in terms of signal processing. When aligned properly, however, the resulting data- sets are comparable to those obtained by common gene expression analysis, enabling the use of existing software packages. Although one can never be entirely sure that all potential sources of systematic bias related to serum peptidomics have been addressed, it has been possible to validate a prostate cancer-specific peptide signature obtained in a recent study using the optimized methodology (9).

By correlating identified proteolytic patterns with several disease groups and controls, it has also been shown that exoprotease activities, superimposed on the ex vivo coagula-
tion and complement degradation pathways, contribute to generation of not only cancer-specific but also “cancer type”-specific serum peptides (9). None of the signature peptides appear to be derived from cancer tissues, implying that different tumor types secrete and/or shed distinct proteases that through their catalytic activity may generate unique serum peptide profiles. The small number of blood proteins that were the source of nearly all the peptides in the prostate, bladder, and breast cancer signatures are therefore not bona fide biomarkers but appear to serve as an endogenous substrate pool for tumor-derived proteases. There was also no relationship between the precursor substrate concentrations and the MS ion intensities of many of the degradation products. For instance, highly abundant serum proteins such as albumin and immunoglobulins were not represented. It therefore appeared that a direct link exists between peptide marker profiles of disease and differential protease activities and that the patterns may have clinical utility as surrogate markers for detection and classification of cancer.

Although critics of serum peptidomics have often pointed at demographics as a potential source of bias, to our knowledge this has never been systematically investigated. We therefore wanted to address whether parameters such as gender and age influence peptidome profiles as obtained using our mass spectrometry-based serum peptide profiling platform. To this end, we collected blood samples from healthy individuals of both genders between the ages of 20 and 80 years and analyzed them using our serum peptidomics methodology. Furthermore this study was specifically intended to verify whether age and gender effects can mask disease effects in serum peptidomics analyses and/or whether differences in peptidome profiles thought to be associated with a specific disease may instead represent demographic bias. The study group was therefore designed to comprise equal numbers of men and women and equal numbers of cancer patients and age-matched healthy individuals. Thyroid cancer, the most common endocrine malignancy, was chosen for this investigation as it is the focus of a major proteomics initiative at our institution.

Since 1973, the incidence of thyroid cancer has increased nearly 50%, and the American Cancer Society projects 30,000 new cases in 2006. There are over 300,000 thyroid cancer survivors in the United States who are under surveillance. Currently there is not a single test that is perfectly capable to predict the presence or absence of thyroid cancer before surgery. The current tumor marker, thyroglobulin, is secreted by all normal thyroid follicular cells and well differentiated thyroid carcinomas. Several thyroglobulin assays are available, but there are several assay limitations that reduce the diagnostic accuracy in predicting residual cancer (10). It has great utility in monitoring patients who have had their entire thyroid removed but makes no distinction between the presence of benign or malignant tissue prior to or after surgery. Patients must therefore undergo fine needle aspiration of any suspicious thyroid nodules and several diagnostic modalities such as ultrasound, radiiodine scanning, and positron emission tomography/computed tomography scanning to determine the existence of metastatic disease. Serum peptidome profiling has the potential to replace presurgical diagnostic modalities as the gold standard for use in thyroid cancer diagnosis and follow-up surveillance testing. It also has the potential to predict which patients harbor residual or recurrent disease and may provide a likelihood prediction of which lesions respond best to radioactive iodine, chemotherapy, or external beam radiotherapy.

Here we report that extensive MALDI-TOF MS and data analysis of serum peptidomes of 200 healthy men and women and of 60 metastatic thyroid carcinoma patients indicated only negligible contributions of both age and gender to the patterns except that healthy men and women under 35 years, but not older individuals, could be distinguished with better than chance accuracy. Through variability analysis of either demographic or disease (versus control) groups, we ruled out demographic bias. A 12-member peptide ion thyroid cancer signature was developed and enabled classification of an independent validation set with 95% sensitivity and 95% specificity. One of the signature peptides was dehydro-Ala3-fibrinopeptide A. As we expand our studies to include hundreds of thyroid cancer patients, the peptide signature will likely be adjusted, further validated, and evaluated in a clinical setting.

EXPERIMENTAL PROCEDURES

Materials

Acetonitrile was obtained from Burdick and Jackson (Muskegon, WI), TFA was from Pierce, and premade α-cyano-4-hydroxycinnamic acid matrix solution was from Agilent (Palo Alto, CA). Human serum (catalog number S-7023, lot 034K937) was obtained from Sigma. Serum peptide processing was done in 0.2-ml polypropylene tubes (8 × 12-tube TempPlate II) from USA Scientific (Ocala, FL).

Serum Samples

Blood samples from volunteer subjects with no known malignancies (Supplemental Table 1) and from patients diagnosed with metastatic thyroid carcinoma were collected following our standard clinical protocol (8) and after obtaining informed consent. Details on patient age, gender, and pathologic diagnosis are given in Supplemental Table 2. All collections were approved by the Memorial Sloan-Kettering Cancer Center Institutional Review and Privacy Board. Blood samples were collected in 8.5-ml, BD Vacutainer SST “tiger-top” tubes (BD Biosciences catalog number 367988), allowed to clot at room temperature for 1 h, and centrifuged at 1400–2000 relative centrifugal force for 10 min at room temperature (8). Sera (upper phase) were transferred to four 4-ml cryovials (Fisher catalog number 0566966), with ~1 ml serum in each, and stored frozen at –80 °C until further use. Upon arrival at the MS laboratory, the cryovials (source vials) were barcoded using the Memorial Sloan-Kettering Cancer Center Clinical Proteomics laboratory information management system (see below) and a 24M barcode printer (Zebra Technologies, Vernon Hills, IL) (8). One cryovial of each sample was thawed on ice and used to generate nine smaller aliquots (50 μl each) in micro-Eppendorf tubes that were also barcoded and stored at –80 °C in
barcoded freezer boxes until further use. Each serum sample had therefore been frozen and thawed twice before it was subjected to solid-phase peptide extraction and MS.

**Automated, Solid-phase Peptide Extraction**

Serum peptide profiling was done as described previously (7, 8). Peptides were captured and concentrated using SMART-C8/K super-paramagnetic, silica-based particles bearing C8 reversed-phase ligands (Chemicell, Berlin, Germany). All analyses were performed in a 96-well format using the same batch of C8 magnetic particles, in 0.2-ml polypropylene tubes, using a Genesis Freedom 100 (Tecan) liquid handling work station. This system automates all 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. A computer randomization program is used to position case and control samples for both solid-phase extraction and mass spectrometry.

**Mass Spectrometry**

Peptide profiles were analyzed with an Autoflex MALDI-TOF mass spectrometer (Bruker, Billerica, MA) as described previously (7). Separate spectra were obtained for two restricted m/z ranges, corresponding to polyepitides with molecular mass of 0.7–4 kDa ("<=4 kDa") and 4–15 kDa (">=4 kDa") (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 spot. The irradiation program was automated using the "AutoXecute" function of the instrument. Spectra were acquired in linear mode geometry under 20 kV (18.6 kV during delayed extraction) of ion accelerating and 1.3 kV multiplier potentials and with gating of mass ions <=400 m/z (<=4-kDa segment) or >=3,000 m/z (>=4-kDa segment). Delayed extraction was maintained for 80 (<=4 kDa) or 50 ns (>=4 kDa) to give time lag focusing after each laser shot. Peptide samples were always mixed with 2 volumes of premade α-cyano-4-hydroxycinnamic acid matrix solution (Agilent), deposited onto the stainless steel target surface in every other column of the 384-spot layout, and allowed to dry at room temperature. A weekly performance test using commercial human reference serum (Sigma catalog number S-7023, lot 034K9397) was done, and the effective laser energy delivered to the target was adjusted as necessary (8).

**Assigning Peptide Sequences**

Peptides selected on the basis of statistical differences in ion intensity between cancer and control groups were analyzed by MALDI-TOF/TOF tandem mass spectrometry using an UltraFlex TOF/TOF instrument (Bruker) operated in "LIFT" mode. The monoisotopic masses were first assigned by one-dimensional reflectron-TOF MS in the presence of three peptide calibrants as described previously (9). Spectra were obtained by averaging multiple signals; laser irradiance and number of acquisitions (typically 100–150) were operator-adjusted to yield maximal peak deflections derived from the digitizer in real time. Monoisotopic masses were assigned for all selected and other prominent peaks after visual inspection, and the low and high end internal standards were used for recalibration. The pass/fail criterion for recalibration is a correct assignment of an m/z value for the "middle" calibrant with a mass accuracy equal or better than 12 ppm.

Alternatively a QSTAR XL Hybrid quadrupole (Q) time-of-flight mass spectrometer (Applied Biosystems/MDS Sciex) equipped with an o-MALDI ion source was used for both duplicate and additional tandem MS analyses. By selecting precursor ions of interest in "Q1" (operated in the mass filter mode), mass measurements of fragment ions could be obtained in the TOF detector following CID in "Q2." Typically a mass window of 3 Da was selected to transmit the entire isotopic envelope of the precursor ion species. Collision energy was operator-adjusted to yield maximum number and intensities of the fragment ions.

Fragment ion spectra resulting from TOF/TOF and Q-TOF analyses (300–1,000 acquisitions averaged per spectrum) were taken to search a "non-redundant" human database (NCBInr; release date, May 20, 2005; 134,668 entries; National Center for Biotechnology Information, Bethesda, MD) using the Mascot MS/MS ion search program, version 2.0.4 for Windows (Matrix Science Ltd., London, UK) with the following search parameters: monoisotopic precursor mass tolerance of 40 ppm, fragment mass mass tolerance of 0.5 Da, and without a specified protease cleavage site. Mascot "mowse" scores greater than 35 were considered significant. Any identification thus obtained was verified independently by two different people by comparing the computer-generated fragment ion series of the predicted peptide with the experimental MS/MS data. Some sequence identifications had below threshold scores but could nonetheless be unequivocally assigned as the precursor ion mass and selected fragment ion masses (b or y) matched a particular peptide in a nested set of sequences (9) (Supplemental Table 3), taking also into account that the limited fragmentation patterns were in agreement with established rules of preferential peptide bond cleavage (e.g. Pro-directed fragmentation).

**Signal Processing**

The custom signal processing software used in this study has been extensively described in an earlier report (8). Data are stored with a naming convention that allows each sample to be associated with its calibrant. The spectra are then converted from binary format to ASCII files and processed in MATLAB with a custom script, "Ocealign," that uses the "Qpeaks" program (Spectrum Square Associates, Ithaca, NY) for smoothing, baseline subtraction, and peak labeling. The singletwidth parameter was set to ~400 for the lower mass range and ~200 for the upper mass range, thereby specifying the resolution, (m/z)/Δ(m/z), for processing. Peak information was used automatically by Qpeaks in setting the parameters for smoothing, baseline subtraction, and binning. The noise statistics were assumed to be "Normal."

Following parameter selection, a setup file was created, and Ocealign queried this file to obtain a list of directories for processing. All data files in all listed directories were aligned with each other during a single processing run. For each directory, singletwidth information was provided in the setup file along with parameters controlling calibration, peak labeling sensitivity, alignment, etc. The files containing the polypeptide standards were first calibrated, the centroid positions of peaks were then obtained from the peak table and compared with the known polypeptide peak positions, and a quadratic calibration equation for correcting the measured masses in each calibration file was created. Ocealign then creates a reference file to which all sample spectra will later be aligned, then loads all other sample files, calibrates them, and adds the intensities to the intensity of the reference file to create the average of all the sample files. The base line was subtracted and normalized to unit size by dividing each intensity value by the total ion count, and a scaling factor was added by multiplying each intensity value by a user-selected number (e.g. 10^4). A peak table, smoothed curve, and a base line were then created, and the spectrum was taken for alignment. A custom alignment algorithm, "Entropycal," was then used to align sample data files to a reference file using a minimum entropy algorithm by taking unsmoothed ("raw"), base line-corrected data (11). The alignment was performed in two steps. At each relative position n, the Shannon entropy of the sum of the two files was computed, and the optimal alignment occurred at the shift that produces the lower value. The

**Thyroid Carcinoma Serum Peptidome Signatures**

1842 Molecular & Cellular Proteomics 5.10
smoothed spectrum was then updated to reflect the aligned \(m/z\) values, and the peak table was updated. The peak lists were 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 was created for further statistical analysis.

**MATLAB Software Tools**

Three software modules, developed in MATLAB, were used for visualization and signal processing of the spectra (8). (i) Signal Processing & Preview (SPP), a graphical viewer for spectra in ASCII format, allows plotting raw and processed spectra side-by-side to review the outcome of signal processing. Furthermore, parameters of Qpeaks (the signal processing software) can be adjusted. (ii) Mass Spectra Viewer (MSV), a visual interface for processed spectral data, plots spectra as \(x-y\) curves (mass versus magnitude) for examining the signatures of several groups of samples. MSV supports regular browsing functions such as scroll, zoom, highlighting, etc. (iii) Heat-Map (HM) displays spectra as two-dimensional heat maps in which the magnitude of the peaks are color-coded on a continuous scale. In addition to browsing functions such as zoom and scroll, the rank of \(x\)- and \(y\)-position coordinates can be reorganized without the constraints of statistical correlation that are enforced by most Heat-Map commercial software packages.

**Statistical Analysis**

The spreadsheet ("peak list"); see "Signal Processing"), containing "binned" data from spectra obtained for all samples of cancer patients and healthy subjects (260 samples total; 598 \(m/z\) values with normalized intensities for each sample; >155,000 data points), were imported into the GeneSpring program (Agilent). Different "experiments" were created in GeneSpring to represent the masses. No normalizations were applied to the experiment because the masses were normalized by the database that binned them. In the "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.

**Class Comparison**—The 598 \(m/z\) values were subjected to average-linkage hierarchical clustering using standard correlation (also known as "Pearson correlation around zero") as a distance metric (GeneSpring program). Peaks were organized by creating mock phylogenetic trees ("dendrograms"). Trees were then displayed with the samples along the \(x\) axis and the masses along the \(y\) axis. Principal component analysis (PCA) on samples was also done using the GeneSpring program.

**Feature Selection**—Once the experiments were created, the \(m/z\) values ("peaks") were filtered by using non-parametric tests: Mann-Whitney \(U\) test (for binary comparisons) and Kruskal-Wallis test (for multiclass comparisons) on each of the 598 features. Significance levels were adjusted to account for the consequences of multiple testing using the method of Benjamini and Hochberg (12) that limits the false discovery rate (FDR). These tests are meant to find peaks that show statistically significant differences between the clinical groups studied.

**Class Prediction**—Support vector machine (SVM) and K-nearest neighbor (K-NN) analyses were done using the Class Prediction Tool in GeneSpring. Leave-one-out cross-validation (LOOCV) experiments were done using both SVM and K-NN modeling on all 598 peptides.

In K-NN modeling the number of neighbors was set to 3, 5, 7, or 10 with a \(p\) value decision cutoff of 1. The kernels used in SVM modeling were: linear, polynomial (order 2), polynomial (order 3), and radial. The class prediction strategy to optimize the thyroid cancer peptide signature was done as follows. Several models using K-NN and SVM algorithms were built using the classification error rate in the cross-validation of the training set \((n=80)\) as the criterion for parameter optimization. In these analyses, different combinations of peptides selected using the Mann-Whitney \(U\) test at different adjusted \(p\) value cutoffs were used to build the models. The best models (the ones giving the smaller classification error rate in the cross-validation of the training set) were then tested in the validation set \((n=40)\). In addition to different combinations of peptides selected using the Mann-Whitney \(U\) test, random combinations of peptides were tested using our class prediction models with the validation set. Five different random combinations of 17, 12, and five features were generated out of the 598 features and tested with the validation set. Their classifications rates were averaged (five different experiments) and compared with the classifications rates of the true combinations of features obtained during feature selection.

**RESULTS**

**Age- and Gender-associated Serum Peptidome Variability Is Minimal**—To evaluate the possible effects of major demographic parameters such as gender and age on serum peptide profiling, we collected blood samples from 200 healthy volunteers and prepared the corresponding sera using a single, standard clinical protocol. Volunteers were recruited for this study to result in three, roughly equal size age groups: 20–35 years old (37 women and 33 men), 35–50 years old (33 women and 34 men), and over 50 years old (39 women and 24 men) (see Figs. 1 and 2, left panel). Gender and precise age of each volunteer are listed in Supplemental Table 1. Post-collection sample handling was uniform, involving two freeze-thaw cycles to accomplish initial storage and subsequent aliquoting for peptide extraction and MS analysis (8). All samples were processed fully automatically (i.e., peptides were extracted on magnetic beads coated with C18 phase, washed, eluted, mixed with matrix, and deposited on the MALDI target plate), as a single batch, and using a customized robot liquid handler followed within 1 h by automated MALDI-TOF mass spectrometric analysis (see “Experimental Procedures”). System reproducibility was always first verified on the same day by analysis, computer alignment, and visual comparison of 12 reference serum samples/spectra as described previously (8). Samples were randomly distributed during processing and analysis. Processed spectra (see “Experimental Procedures”) were aligned using the custom Entropycal algorithm (8), and a total of 598 distinct \(m/z\) values, or peaks, were resolved in the 700–8,000-Da range. A spreadsheet (peak list), containing the normalized intensities (i.e. signal intensities, after baseline subtraction, were divided by the total ion current of the corresponding spectrum and multiplied by a scaling factor of \(1 \times 10^7\)), of all 598 peaks for each of the 200 samples was then taken for unsupervised analysis.

Overall peptidome profiles of all samples were analyzed by average-linkage hierarchical clustering using standard corre-
The algorithm orders peptides on one axis and samples on the other based on similarity of their profiles. No clear correlation was found between peptidome profiles and demographic parameters when the resulting dendrogram was colored based on either age groups or gender as shown, respectively, in Fig. 3, left top and bottom panels; i.e. the colors are largely intermingled. PCA, also an unsupervised analysis, did not show any meaningful sample grouping either. PCA helps to reduce the high dimensionality of a dataset by calculating the contribution of the signal of each peptide to the variance across the sample set and assigning it to a component. Usually most of the variance of a dataset can be explained using three components that can then be visualized in a three-dimensional scatter plot. In this case, the three components shown in Fig. 3 (top and bottom panels on the right) explain 91.28% of the variance of our dataset.

Next we carried out a supervised analysis to identify peptides that showed significant differences in MS ion intensities between the various age groups or between men and women. To this end, a Mann-Whitney U test was performed on each of the 598 peptides and using either all 200 samples or well defined subsets thereof (Table I); significance levels were adjusted to account for the consequences of multiple hypoth-
esis testing using the method of Benjamini and Hochberg (12) that limits the FDR. Again neither age or gender appeared to affect the peptide profiles because the number of differential peptides (“features”) selected at a $p$ value of less than 0.05 (i.e. 5% by chance) were ≤2.5% of the total number; no features were found for every two- or three-way comparison with a threshold of $p < 0.00001$ (Table I). Conversely more than 25% of all peptides passed the more stringent threshold (i.e. $p < 0.00001$ with FDR adjustment) in a comparable discriminant analysis between serum peptide profiles of thyroid cancer patients and controls. These features were then successfully used in the validation of an external sample set (9).

We subsequently carried out class prediction analyses using LOOCV with either SVM or K-NN modeling. All peptides (598) were used, and the same two- or three-way comparisons as for the feature selections were done. As shown in Table I, the ability of this 598-feature set to correctly predict the different demographic subgroups by machine learning was only just better than chance, if at all, with one as yet unconfirmed exception. In a two-way comparison, men and women under 35 years of age could be assigned to the correct gender with 70% accuracy (by SVM and K-NN). No such result was obtained for men versus women older than 35 years or for any of the age group comparisons, be it of mixed or separate gender. We identified 15 peptides that somewhat distinguish $p < 0.04$ with FDR adjustment) males and females in the younger group. Only six of these peptides ($m/z = 1,278, 1,351, 1,752, 1,787, 2,115, and 2,553$) corresponded to members of previously established cancer serum peptidome marker patterns (9); the rest were unknown.

### Table I

**Serum peptidome patterns and class prediction based on age and gender**

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Men</th>
<th>Women</th>
<th>All ages</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>m/w</td>
<td>33</td>
<td>52.6</td>
<td>50.8</td>
</tr>
<tr>
<td>35-50</td>
<td>m/w</td>
<td>34</td>
<td>53.4</td>
<td>55.9</td>
</tr>
<tr>
<td>&gt;50</td>
<td>m/w</td>
<td>34</td>
<td>54.1</td>
<td>55.9</td>
</tr>
<tr>
<td>&lt;35</td>
<td>m/w</td>
<td>34</td>
<td>54.9</td>
<td>55.9</td>
</tr>
<tr>
<td>35-50</td>
<td>m/w</td>
<td>59</td>
<td>66.7</td>
<td>64.1</td>
</tr>
<tr>
<td>&gt;50</td>
<td>m/w</td>
<td>61</td>
<td>71.4</td>
<td>56.4</td>
</tr>
<tr>
<td>&lt;35</td>
<td>m/w</td>
<td>56</td>
<td>71.4</td>
<td>56.4</td>
</tr>
<tr>
<td>35-50</td>
<td>m/w</td>
<td>64.3</td>
<td>71.4</td>
<td>54.5</td>
</tr>
<tr>
<td>&gt;50</td>
<td>m/w</td>
<td>49.3</td>
<td>71.4</td>
<td>54.5</td>
</tr>
</tbody>
</table>

**By chance**

<table>
<thead>
<tr>
<th>SVM (LOOCV)</th>
<th>K-NN (LOOCV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear kernel</td>
<td>n = 3</td>
</tr>
<tr>
<td>Polynomial 2</td>
<td>39.5</td>
</tr>
<tr>
<td>Polynomial 3</td>
<td>37.5</td>
</tr>
<tr>
<td>Radial</td>
<td>39.5</td>
</tr>
<tr>
<td>n = 5</td>
<td>61.7</td>
</tr>
<tr>
<td>n = 7</td>
<td>56.4</td>
</tr>
<tr>
<td>n = 10</td>
<td>51</td>
</tr>
<tr>
<td>By chance</td>
<td>33.3</td>
</tr>
</tbody>
</table>

**TABLE I**

**Serum peptidome patterns and class prediction based on age and gender**

<35, 20–35 years old; 35–50, 35–50 years old; >50, over 50 years old; m, men; w, women; ANOVA, analysis of variance.
Thyroid Carcinoma Serum Peptidome Signatures

Cancer Is a Major Source of Variability in Serum Peptidome Profiles—As part of an ongoing, related project at our host institution, we have also collected and analyzed sera from 60 patients who had clear evidence for residual metastatic thyroid carcinoma (30 men and 30 women aged between 15 and 86) (Figs. 1 and 2, right panel). Specimens are linked to database records (listed in Supplemental Table 2) but were anonymized and stripped of any patient identifiers to meet Health Insurance Portability and Accountability Act guidelines. Blood collection, sera preparation, storage and handling, automated peptide solid-phase extraction, and mass spectrometry were done exactly as for the 200 controls. In fact, all 260 serum samples were analyzed simultaneously but in a positionally randomized manner (i.e. positioning in the microtiter plate wells and on the MALDI target plates was determined by a computer randomization program).

When the 260 peptidome profiles were analyzed by average-linkage hierarchical clustering using all 598 peptides, two roughly separated clusters emerged. The bigger cluster (Fig. 4A, left side of the dendrogram) contained 191 of the 200 healthy volunteer samples (colored in red), whereas 56 of the 60 thyroid cancer patient samples (colored in blue) were contained in the cluster on the right. This observation is in clear contrast with the scattering of gender and age groups in the dendrograms shown in Fig. 3. PCA revealed a similar grouping with the healthy controls and the metastatic cancers in largely separated clusters (Fig. 4B). Taken together, the clusters in Figs. 3 and 4 suggest that thyroid cancer (i.e. patients harboring metastatic lesions) is by far the major source of variance in our 260-sample serum peptidome dataset, resulting in near completely separable groups except for four outlying patient samples (573a, 655a, 758a, and 764a are highlighted in Supplemental Table 2) and nine outlying controls (four males, 23, 24, 33, and 35 years old, and five females, 27, 50, 51, 51, and 57 years old; highlighted in Supplemental Table 1). We do not have a ready explanation for the outliers.

In a third and final analysis, we examined the variance (analysis of variance) for each of the 598 peptides for each of the three study cohort parameters. This test calculates the association strength between the peptidome profiles and the parameters under study (see “Experimental Procedures”). A Mann-Whitney U test (to compare both genders or to compare healthy controls versus disease) or Kruskal-Wallis test (to compare the three age categories) was performed, and a p value was calculated for each peptide in each comparison. Using a p < 0.05 cutoff, 110, six, and three differential peptides were retrieved from the (i) cancer versus control, (ii) gender, and (iii) age group comparisons, respectively; the smallest p value obtained for each these comparisons was, respectively, 9.97 × 10^{-6}, 3.99 × 10^{-5}, and 1.81 × 10^{-4}. The negative logs of all p values were then summed to create a measure termed “association strength” (i.e. when smaller p values are generated for more significant parameters, association strength will increase), which was 3,137, 779, and 638, respectively. Overall our analyses revealed that a much larger number of discriminating peptides more strongly associated with disease than with any particular age group or gender.

MALDI-TOF MS-based, Peptide Ion Patterns Differentiate Metastatic Thyroid Carcinoma Patient Sera from Matched Controls—To further exclude any uncontrolled effects of gender or age, no matter how little, on serum biomarker pattern discovery studies, we selected a set of 60 gender- and age-matched controls (30 men and 30 women) from the earlier described 200-sample set and compared those with 60 samples obtained from 30 men and 30 women with metastatic thyroid carcinomas (see Fig. 1). This latest “120-sample” set was then taken for all the work described hereafter to identify peptides with thyroid carcinoma biomarker potential. First training and validation sets were randomly created (Fig. 1; age distribution in Fig. 5). Unsupervised analysis of the training set (40 patients and 40 matched controls) peptide profiles by average-linkage hierarchical clustering allowed satisfactory segregation of the thyroid cancer and control samples without any prior statistical test to select the most discriminatory peaks (i.e. all 598 peptides were used) (Fig. 6A). On the other hand, no correlation of the peptidome profiles with gender was observed (Fig. 6B). PCA revealed the same grouping, or absence thereof, with control and thyroid carcinoma samples in different clusters but with genders intermingled (Fig. 6, A and B).

We then performed a supervised analysis (Mann-Whitney U test), using all 80 samples and 598 peptides in the training set, to identify peptides that showed statistically significant differences between thyroid cancer and control. Twenty peptides were found after applying an FDR-adjusted p value cutoff of 1 × 10^{-5}. This number was further reduced to 17 by applying...
an arbitrary threshold to the median ion intensities of each individual peak within each sample cohort exactly as done in a previous study (9). The threshold was set high enough to select robust peaks in the spectra with intensities that would permit MALDI MS/MS-based tandem mass spectrometric sequencing and to exclude closely positioned neighboring peaks or “shoulders.” Twelve of those 17 selected ion peaks corresponded to peptides with molecular mass \( <2,000 \text{ Da} \) (assuming \( z = 1 \)), four were between 2,000 and 4,000 Da, and only one peptide had a mass \( >4,000 \text{ Da} \) (Figs. 7 and 8). To verify those findings, each of the 17 selected \( m/z \) peaks was also visually inspected in color-coded (cancer, blue; controls, red) overlays of all 80 training set spectra (Fig. 7). As can be readily observed, \( m/z \) peak intensities may differ dramatically between the two groups; sometimes intensities were consistently higher in all thyroid cancer specimens (seven of 17), and sometimes intensities were higher in the controls (10 cases).

The results, when represented in the form of heat maps, as shown in Fig. 8, indicated that data reduction by 97% (from 598 to 17 peptides) did not adversely affect the separation of the clinical groups. Only one of the peptides that we had found to somewhat distinguish (\( m/z = 1,350.64 \); Mann-Whitney \( p = 0.032 \) with FDR adjustment) young males and females in the healthy control group was also part of the thyroid cancer 17-peptide signature but with much stronger capacity to classify (Mann-Whitney \( p = 3.5 \times 10^{-24} \) with FDR adjustment).

To verify whether the results were not affected by irrepro-
each individual analysis independently. Triplicate runs were then compared, and peaks were automatically selected and labeled for each analysis. The peak list (598 ± 500 “units”) was subjected to a Mann-Whitney U test for each individual cancer versus control. Only peaks with adjusted p values <0.00001 were passed through a second filter (median peak intensity, >500 “units”); a peak was selected if it passed the intensity threshold in either cancer or control. The heat map compares the 17 selected features (the thyroid cancer signature) for each sample in the study set. Columns represent samples (per group); rows are m/z peaks as indicated by the monoisotopic mass (on the left) and peptide sequence (on the right) for each. Serum peptide sequences highlighted in pink are newly identified in this study. The longest peptide (i.e. ending in *** ) is shortened as listed here: the complete sequence (54 amino acids) is SSSYSKQFTSSTSNSRAGDSTFESKSYKMADEAGSEADHE-GTHSTKRGHAKSRPV. Thyroid MET, patients with metastatic thyroid carcinomas.

Fig. 8. Feature selection and comparative analysis of serum peptide profiling data derived from thyroid cancer patients and healthy controls. The peak list (598 × 80) was subjected to a Mann-Whitney U test for each individual cancer versus control. Only peaks with adjusted p values <0.00001 were passed through a second filter (median peak intensity, >500 “units”); a peak was selected if it passed the intensity threshold in either cancer or control. The heat map compares the 17 selected features (the thyroid cancer signature) for each sample in the study set. Columns represent samples (per group); rows are m/z peaks as indicated by the monoisotopic mass (on the left) and peptide sequence (on the right) for each. Serum peptide sequences highlighted in pink are newly identified in this study. The longest peptide (i.e. ending in *** ) is shortened as listed here: the complete sequence (54 amino acids) is SSSYSKQFTSSTSNSRAGDSTFESKSYKMADEAGSEADHE-GTHSTKRGHAKSRPV. Thyroid MET, patients with metastatic thyroid carcinomas.
None of the samples in the validation set had been included in the earlier supervised analysis, therefore allowing for the estimation of true predictive accuracy. Either all 598 peptide ions or the 17-peptide signature was used for comparison of four separate groups (Fig. 9A: cancer/training, blue; control/training, red; cancer/validation, yellow; control/validation, light blue) by PCA (Fig. 7B). In this unsupervised analysis, samples from thyroid cancer sets 1 and 2 were relatively well separated from the control sets 1 and 2.

Using the training set, a class prediction strategy was then devised to further optimize the thyroid cancer peptide signature. To this end, several models were built using K-NN and SVM machine learning algorithms, and the classification error rate in the cross-validation of the training set (n = 80) was used as a criterion for parameter optimization. Specifically we first tested the complete 598-peptide pattern as well as three smaller sets that had been selected using the Mann-Whitney U test with FDR-adjusted p value cutoffs of 1 x 10^{-5} (the original 17-peptide signature; Figs. 7 and 8), 1 x 10^{-7} (12 peptides), and 1 x 10^{-10} (five peptides). We then tested the best models, i.e. giving the smallest classification error rates in the cross-validation of the training set, for class prediction of the validation set (n = 40). The results shown in Fig. 10 indicate excellent specificity (100%) but poor sensitivity (\sim 70%) of the full 598-peptide pattern to classify thyroid cancer samples and matched controls using K-NN with three, five, seven, or ten neighbors. By contrast, the smaller peptide signatures could classify samples with sensitivities and specificities of \sim 90% or better. The best overall results were obtained for the 12-peptide signature (containing the two newly identified peptides, including dehydro-Ala-FPA), yielding 95% sensitivity and 95% specificity (the binomial confidence interval at 95% was 75.1–99.9% for 19 correct predictions of 20 cancer cases or 20 controls) (Fig. 10). To substantiate these results, we tested random selections of features for class prediction accuracy of the validation set using the same models above. Random combinations of 17, 12, and five peptides were generated five times each, and the classification error rates were averaged and then also tested for class prediction accuracy of the validation set using the same models.
DISCUSSION

Genomic diversity is generally thought to be the major biological factor determining phenotypic differences. However, variations in gene expression and post-transcriptional events also exist and likely contribute to diversity throughout phylogeny (17–20). Variation in gene expression appears to be at least in part genetically determined, and its regulation is complex. DNA microarray technology is commonly used for mRNA expression profiling and to identify individual genes that exhibit changing levels across physiologic conditions and/or pathological states. In doing so, a large number of expression patterns have been measured, thereby defining differential patterns between tissues, individuals, and species (18, 21–23). Effects of aging and gender on gene expression have also been documented (19, 24–26). The use of model systems including yeast, worms, flies, and mice as well as studies of human progenies and cellular senescence have furthermore identified epigenetic events believed to contribute to the “aging” phenotype (27). These include the effects of oxidative damage associated with cellular metabolism, genome instabilities such as telomere shortening, mitochondrial mutations, mitotic machinery errors, and chromosomal pathologies (27, 28). Consequently age and perhaps other demographic factors may affect the composition and relative abundance of molecular patterns (“signatures”) that are increasingly being evaluated by clinical investigators for possible diagnostic and prognostic purposes. In this regard proteins and peptides are among the most frequently measured molecules certainly in those easy to obtain biological fluids such as blood or urine.

We have developed and applied a unique serum profiling platform (7) to study a large cohort of identically collected and processed samples from 200 healthy men and women, ages 20–80, and from 60 patients (30 men and 30 women) with metastatic thyroid carcinoma. Extensive MALDI-TOF MS and data analysis suggested only negligible contributions of both age and gender to the serum peptidome pattern. Feature selection and statistical analysis further indicated that possible distinguishing characteristics were generally no better than what could be expected by chance. For instance, leave-one-out class predictions for gender or age group comparisons gave around 40–60% accuracy with one minor exception. In the age group under 35 years, but not in any other age brackets, healthy men and women could be distinguished with ∼70% accuracy by LOOCV based on the serum peptide patterns. Although this is by no means an accurate classification rate, it is likely better than chance in a study cohort of 70 individuals (33 men and 37 women; ∼50 correctly classified by LOOCV). We can only speculate at this time what the underlying reason(s) might be for this observation. All the same, the effect is rather subtle and, considering the more advanced age of most cancer patients, unlikely to interfere with peptidomics analysis of cancer sera.

We nonetheless verified whether differences in serum peptidome profiles thought to be associated with cancer could instead represent demographic bias. A study group was therefore assembled to comprise equal numbers of men and women and equal numbers of metastatic thyroid cancer patients and age-matched healthy controls. Unsupervised hierarchical clustering resulted in a fairly clear-cut separation of cancer and controls, whereas gender distribution was basically random. In addition, a much larger number of discriminating peptides more strongly associated with disease than with any particular age group or gender. Using statistical and other filtering methods and a class prediction optimization strategy based on LOOCV and machine learning methods, a 12-peptide thyroid cancer signature was then obtained and enabled classification of a totally independent validation set with 95% sensitivity and 95% specificity. Interestingly 10 of these peptides had previously been assigned to signature patterns of other solid tumors (e.g. breast, prostate, and bladder cancer) (9). Unfortunately the use of different blood collection tubes in the current study and in those earlier studies precluded meaningful comparisons or multiclass predictions of the thyroid cancer signature.

One of the two unique, newly identified peptides in the thyroid cancer signature is dehydro-Ala-FPA, which is most likely derived by β-elimination from the phospho-Ser residue (13) known to naturally occur at that position in about 20–30% of FPA and fibrinogen molecules (14–16). Curiously we have never observed phospho-FPA in either cancer patient or control samples, a fact that may actually be related to poor MALDI in positive mode (29). Therefore, dehydro-Ala-FPA could simply be a surrogate for phospho-FPA in MALDI-TOF MS-based serum profiling. We do not know at this point whether the conversion occurred in serum or during sample processing or mass spectrometry (30). It is puzzling, however, why β-elimination, generally thought to be induced by alkaline conditions (13), could have occurred at the low pH that is maintained throughout our serum processing protocol. It is also of note that phospho-Ser3-FPA has previously been detected at elevated levels in sera of ovarian cancer patients (31). By contrast, we found dehydro-Ala-FPA at significantly lower concentrations in thyroid cancer patient sera relative to the controls.

In summary, the serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age and appear to have diagnostic potential with high sensitivity and specificity to classify samples in an independent validation set. We believe as we expand the current study to include hundreds more thyroid cancer patients that the peptide signature can be appropriately adjusted, optimized, and further validated. It will then be evaluated in a clinical setting either used independently or in combination with existing diagnostic procedures such as fine needle aspiration biopsy or serum thyroglobulin measurements.
Furthermore our approach can be generalized for many diagnostic and predictive purposes, as an in vitro phenotypic readout of catalytic and metabolic activities in body fluids or tissues, utilizing either endogenous substrates or measured quantities of externally added, isotopically labeled substrate analogs followed by quantitative product analysis. Various analytical readouts, product selection schemes, and activity attenuation procedures can be envisioned to provide more, or different, data points and to tailor the process to each specific case of pattern discovery. Alternatively identification and characterization of the protease panels may lead to direct immunoassay-based, quantitative diagnostic tests suitable for a clinical environment.

Acknowledgments—We thank all volunteers for generous donation of blood samples as well as Dr. Lilian Reich and the staff members of Memorial Sloan-Kettering Cancer Center’s blood bank for enthusiasmastic assistance. We are indebted to Dr. Hedyi Erdjument-Bromage for expert help with TOF/TOF MS/MS-based peptide sequencing.

* This work was supported by National Institutes of Health Grants 1 R21 CA1119425, 5 P30 CA08748, and 5 P50 CA92629; the Himanshu Vadi Research Fund; and the Byrne Fund. The costs of publication of this article were defrayed in part by the payment of page charges. §§ To whom correspondence may be addressed: Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021.

+ To whom correspondence may be addressed: Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021.

**§ To whom correspondence may be addressed: Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021.

REFERENCES


---

**Thyroid Carcinoma Serum Peptidome Signatures**

**Molecular & Cellular Proteomics 5.10**
oxidative stress, and apoptosis in mammalian aging. Science 309, 481–484