Detection of breast cancer biomarkers in nipple aspirate fluid by SELDI-TOF and their identification by combined liquid chromatography-tandem mass spectrometry

JIANBO HE[1,2], JEFFREY GORNBEIN[3], DEJUN SHEN[1,2], MING LU[1,2], LEONOR E. ROVAI[1,2], HUNGYI SHAU[1,2], JOHN KATZ[1,2], JULIAN P. WHITELEGGE[4], KYM F. FAULL[4] and HELENA R. CHANG[1,2]

[1]Gonda/UCLA Breast Cancer Research Laboratory, [2]Revlon/UCLA Breast Center, Department of Surgery, [3]Department of Biomathematics, David Geffen School of Medicine, Los Angeles, CA; [4]The Pasarow Mass Spectrometry Laboratory, Department of Psychiatry and Biobehavioral Sciences and the Neuropsychiatric-Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA 90095, USA

Received July 4, 2006; Accepted September 1, 2006

Abstract. Screening mammography is the most effective tool available for breast cancer detection. While screening mammography saves lives, it has intrinsic problems that limit further improvement. We hypothesize that protein biomarkers in nipple aspirate fluid (NAF) may separate the cancer from the non-cancer state, and therefore can be used for breast cancer detection. In this study the proteins in NAF were analyzed by surface-enhanced laser desorption ionization coupled to time-of-flight mass spectrometry (SELDI-TOF) in the m/z 5,000-85,000 range. Two methods were used to normalize spectra. Then differentially expressed signals that separate cancer from non-cancer conditions were selected by two specifically developed statistical algorithms. Proteins of interest were identified by combined liquid chromatography-tandem mass spectrometry. A set of 8 markers were identified which collectively gave 63% sensitivity, 89% specificity and 76% accuracy for distinguishing cancer from non-cancer. Further improvements in the specificity and sensitivity of this strategy could come from the development of methods for more precise quantification of the biomarkers of interest and also from focusing on the low abundant components that are not evident when unfractionated NAF is analyzed directly.

Correspondence to: Dr Helena Chang, Revlon/UCLA Breast Center, 200 UCLA Medical Plaza, Suite B265, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA
E-mail: hchang@mednet.ucla.edu

Abbreviations: AUC, area under the curve; BSA, bovine serum albumin; Da, dalton (one twelfth the mass of carbon-12); HPLC, high performance liquid chromatography; LC/MS, combined liquid chromatography mass spectrometry; LC/MS/MS, combined liquid chromatography tandem mass spectrometry; MALDI, matrix-assisted laser desorption/ionization; μLC/MS/MS, microbore liquid chromatography coupled to tandem mass spectrometry; MS, mass spectrometry; MS/MS, tandem mass spectrometry; NAF, nipple aspirate fluid; SELDI, surface-enhanced laser desorption ionization; SELDI-TOF, surface-enhanced laser desorption ionization-time of flight; TIC, total ion current; TOF, time-of-flight

Key words: breast cancer, biomarker, nipple aspirate fluid, SELDI-TOF, LSMS

Introduction

Despite advances, current treatment remains largely ineffective for late-stage breast cancer. In contrast, the same treatment is successful for early-stage breast cancer (1). Mammography has been shown to be the most effective screening tool for finding breast cancer early and for saving lives. However, mammography has intrinsic limitations that may be difficult to overcome. For example, mammography is ineffective for evaluating dense breasts, and is only marginally useful in young women (2). Biomarker-based laboratory tests may be complementary to the conventional screening methods for early detection of breast cancer. Discovering biomarkers characteristic of breast cancer in body fluid by proteomic technology has become a research focus in recent years.

NAF, first introduced for clinical purposes by Papanicolaou in 1958, contains ductal secretion, cells and cellular contents of the breast ductal-lobular system (3). Although NAF is traditionally used for cytological assessments (4-6), it can also be used to study protein expression patterns by proteomic technologies.

Of the available proteomic techniques, matrix-assisted laser desorption ionization (MALDI), combined liquid chromatography-mass spectrometry (LC/MS) and combined liquid chromatography-tandem mass spectrometry (LC/MS/MS) have generated interest for both displaying a panel of proteins and/or for their identification. These relatively new techniques are now being used to discover biomarkers. A related technology, surface-enhanced laser desorption ionization (SELDI) mass spectrometry, used in conjunction with time-

of-flight mass spectrometry (TOF), is also used to display protein profiles (7-10). An important feature of SELDI-TOF is the unique surface on which the biological sample is captured. These surfaces (so-called ProteinChips®) are covalently modified with one of a number of functionalities. The different types of surfaces include anion and cation exchange (for retaining negatively and positively charged analytes, respectively), metal affinity (for capturing His-tagged proteins, for example), reverse phase (immobilized hydrophobic surface) and normal phase (immobilized hydrophilic surface). The functionalized surfaces are used to capture specific classes of analytes (proteins) and to concentrate them from impure extracts. This process of selectively concentrating desired proteins and peptides circumvents many of the unwieldy steps used in traditional protein analysis (11,12).

A goal of proteomics in cancer research is to perform qualitative and quantitative analysis of all the proteins expressed in body fluids or in tissues. Comparisons made between samples collected from patients with cancer and samples from cancer-free control subjects may reveal unique proteins or changing levels of specific proteins characteristic of the disease. Analysis of the resulting spectra by different types of univariate bioinformatic algorithms, such as peak alignment/clustering, mean/median profile comparisons and multivariate algorithms such as logistic regression, classification trees, neural nets, genetic algorithms, and random forest algorithms (13), enables the application of proteomic data sets to cancer diagnosis. Although they work in different ways, all these algorithms can be used to earmark signals that may be useful for segregating cancer from non-cancer states.

We hypothesize that protein biomarkers existing in NAF may distinguish cancer from the non-cancer state. In the present study, we focus on the comparison of SELDI-TOF spectra of two groups: NAF derived from non-cancerous breasts versus from those with invasive tumor. Two normalization methods, total ion current (TIC) and area under the curve (AUC) are used under the assumption that the peak intensities are quantifiable. We also developed two statistical analysis algorithms for comparing the spectra from the two groups of samples: repeated measure median profile comparison using non-parametric methods and alignment/cluster analysis using the spectral signals.

Materials and methods

Chemicals and reagents. Bovine cytochrome C, myoglobin, and serum albumin (BSA) were obtained from Michrom Biosources (Auburn, CA). Acetonitrile, isopropanol, formic and acetic acids were obtained from Fisher Scientific (Pittsburgh, PA). Sinapinic acid and trifluoroacetic acid were obtained from Sigma-Aldrich (St. Louis, MO). Quartz distilled water (>16 $m\Omega cm^{-1}$) was produced in-house, and all other reagents and solvents were of analytical grade or better.

Collection of NAF. Fluid was obtained from women who consented to a biomarker discovery study approved by the UCLA Institutional Review Board. Upon collection on ice, each NAF sample was divided into aliquots (1 μl) and stored at -80˚C. From women with no known cancer, one breast was sampled from 30 individuals and both breasts were sampled from 8 individuals. From women with invasive breast cancer, one cancerous breast was sampled from 21 individuals and both breasts were sampled (one with no cancer and the other with invasive cancer) from 17 individuals. In total, the study includes analysis of 101 samples from 76 different women, including 63 samples from non-cancerous breasts and 38 samples from breasts with invasive cancer. Patients who previously had cancer removed by surgery were excluded. The data analysis is confined to a comparison of NAF from cancerous and non-cancerous breasts.

SELDI-TOF. NAF was diluted 1:100 in distilled water. One μl of the diluted sample was applied to the wells on NP20 ProteinChips with 1 μl of matrix consisting of saturated solution of sinapinic acid in 70% acetonitrile containing 0.1% trifluoroacetic acid. The SELDI-TOF analysis was performed without washing the spots. Positive ion laser desorption mass spectra were recorded from each dried surface with a linear time-of-flight mass spectrometer (TOF, Protein Biological System II, Ciphergen Biosystems, Inc.), externally calibrated with bovine cytochrome C (12,230.9 Da) and bovine albumin (66,410 Da) by averaging the spectra from 91 laser shots using a constant laser intensity, a deflector setting of 3,000 Da, a constant detector sensitivity and a detection range of 5,000-100,000 Da. The raw data were transferred to the Ciphergen Express Data Manager Software version 3.1 for analysis.

Normalization of the mass spectra. This was performed after subtracting any non-constant baseline using the convex hull algorithm provided by the Ciphergen software (Protein Chip® Software 3.1-Operation Manual 2002, Ciphergen). Two commonly used normalization methods were applied to each spectrum: i) normalization by dividing each signal intensity in the spectrum by the area under the entire spectrum (area under the curve, AUC), and ii) normalization by multiplying each signal intensity by the ratio of the mean intensity (total ion current, TIC) across all spectra divided by the individual mean spectrum intensity. Spectral analyses were limited to the m/z range 5,000-85,000. There are other normalization methods available. In particular, a method often used is internal normalization of each spectrum by dividing each signal intensity by the maximum intensity in the spectrum. This method was not selected because the m/z value at which each maximum occurred varied widely from spectrum to spectrum, implying that this was not a uniform criterion.

Screening analyses. Two different screening methods, median intensity profile and peak cluster/alignment, were performed on the normalized spectra, and the combined results were used to establish a set of potential markers. The median intensity profile method is a non-parametric repeated measure analysis comparing median intensities in one group (cancer) to another group (non-cancer) across pre-determined bins. This method used all signal intensities across the entire m/z range and does not rely on peak identification. Medians were computed bin by bin across all spectra in the same group and compared between the two groups bin by bin as defined below. The bins were formed by first aligning and then sorting

the spectra by m/z, and assigning the first 100 observations to the first bin, the next 100 observations to the second bin etc. Since there is a constant sample size of 100 observations per bin per spectrum, the m/z bin width increases as m/z increases, reflecting the decreasing precision in the data with increasing m/z. The bin width in this study ranged from m/z 7-27, making the bin width <0.2% of the average m/z value in the bin.

The peak cluster/alignment method used only the peaks from each spectrum. Peaks were first identified using the Ciphergen Protein Chip software (Protein Chip Software 3.1-Operation Manual 2002, Ciphergen) at default settings. For each spectrum, the peak intensities and corresponding m/z values were identified and stored. Then a hierarchical clustering of the m/z values was carried out to align peaks from different spectra. Peaks assigned to the same cluster were assumed to have the same true m/z value, differing only because of random error. In this analysis, m/z values that differed by <2% were usually assigned to the same cluster and clusters that did not contain peaks from at least 10% (10/101) of the spectra were omitted. Spectra that did not have a peak in a given cluster were included using their (usually low) observed intensity. After alignment, the median intensities were compared between groups (cancer versus non-cancer), cluster by cluster and expressed as median percent differences and ratios as below.

Screening criteria. For both methods, if A is the median intensity in one group (cancer) and B is the median intensity in the other group (non-cancer), the median difference is A-B, the median percent difference is defined as 100 x (A-B)/[(A+B)/2] and the median ratio is A/B. The non-parametric Wilcoxon's rank sum test was used to compute a p-value for comparisons at each bin or cluster. For either screening method, an m/z value was flagged as a potential marker if: a) the median ratio A/B was <0.8 or >1.2, or b) the median percent difference was >25% in either direction, or c) the non parametric p-value was <0.06. Note that (a) and (b) are almost identical screening criteria.

Multivariate analyses. Using group (cancer or non-cancer) as the outcome and the potential markers from either method from the screening, a stepwise logistic regression analysis was used to identify a final set of markers. Although a classification tree (CART-recursive partitioning) analysis was also carried out, the results were not helpful and are not reported. A cancer score was computed using the marker logit score (the linear combination of the final set of markers that best separated the two groups) and a receiver operator characteristic analysis (ROC) was performed to determine the nominal sensitivity and specificity of the score (and thus the nominal sensitivity and specificity of the combined final markers). The nominal sensitivity and specificity values from the ROC analysis that maximize accuracy are reported, where accuracy is defined as the sum of the sensitivity and specificity divided by 2 [accuracy = (sensitivity + specificity)/2].

Protein fractionation. NAF aliquots (30-100 μl) were mixed with cold 80% acetone (-20˚C, 1 ml), incubated (1 h, -20˚C) then centrifuged (20,000 x g, RT, 2 min). The resulting supernatant was removed. The pellet was re-dissolved in 60 μl of 90% formic acid, and 20 μl was injected onto an HPLC column (PLRP/S, 300 Å, 3 μm, 150x2.1 mm, Polymer Labs) equilibrated in 95% buffer A (0.1% TFA in water) and 5% buffer B (acetonitrile/isopropanol/TFA, 50/50/0.05 v/v). The column was eluted (25 μl/min) with an increasing concentration of buffer B (min/% B: 0/5, 5/5, 30/40, 150/100) at 40˚C. The eluant was passed through a 280-nm UV detector and then a flow splitter. A portion (20-40%) of the eluant was directed to an Ion Spray™ source attached to a triple quadrupole mass spectrometer (Perkin-Elmer Sciex API III$^+$) operating in the MS mode (scan range m/z 600-2,300, step size 0.3 Da, 6.08 sec/scan, orifice ramped between 60-120 V). The remainder of the eluent was collected in 1-min fractions. Selected fractions were also screened by SELDI-TOF and in some instances by MALDI-TOF (Applied Biosystems DE STR) using sinapinic acid matrix.

Protein identification. Aliquots (10 μl) of selected fractions were first treated with dithiothreitol (10 mM in 50 mM ammonium bicarbonate, 15 μl, 1 h, 37˚C), then iodoacetamide (50 mM in 50 mM ammonium bicarbonate, 15 μl, 1 h, 37˚C), then with trypsin (Promega sequencing grade, 12.5 μl, 6 ng/μl in 50 mM ammonium bicarbonate, 3 h, 37˚C). Finally the samples were dried by centrifugal evaporation and re-dissolved in 5 μl of 70% acetic acid before injection onto a microbore polymeric reverse phase HPLC column (PLRP/S, 5 μm, 300 Å, 0.2x150 mm, Michrom Bioresources) equilibrated in water/acetonitrile/formic acid (95/5/0.1, v/v) and eluted (1 μl/min) with an increasing concentration of acetonitrile (min/% acetonitrile: 10/5; 50/40; 65/80; 70/80). The column eluant was directed to a nanospray ion source using coated emitter tips at 3.2 kV attached to an ion trap mass spectrometer (LCQ-DECA, ThermoFinnigan, San Jose, CA). Data-dependent acquisition parameters were: survey scan 400-1500 m/z; zoom scan on dominant ion and product ion scan when multiply charged; two MS/MS scans for each parent ion using dynamic exclusion including exclusion of singly charged parent ions. Data sets were screened against human genomic databases of predicted or known open reading frame translations (Sequest™, ThermoFinnigan, San Jose, CA). Significant matches were examined manually to confirm assignments.

Removal of abundant proteins. An albumin and IgG removal kit (Amershan Biosciences AB, Uppsala, Sweden) was used according to the manufacturer's instructions. Briefly 1 μl of NAF was diluted with 10 μl of water to which was added 100 μl of a slurry containing the antibody-coated beads. The sample was incubated at 4˚C for 30 min, interrupted by gentle mixing every 10 min. Controls included substitution of 10 μl of water for the beads. Samples were then centrifuged (15,000 x g, 3 min, 4˚C) and the upper phase was transferred to a clean tube and 1 μl of each was applied to an NP20 ProteinChip for SELDI-TOF analysis.

A hemoglobin removal kit with Ni-NTA magnetic agarose beads was used according to the manufacturer's instructions (Qiagen, Valencia, CA). The beads were washed 3 times with phosphate-buffered saline (10 mM phosphate, 138 mM sodium chloride, 2.7 mM potassium chloride, pH 7.4). One μl of NAF was diluted with 10 μl of water and added to 10 μl

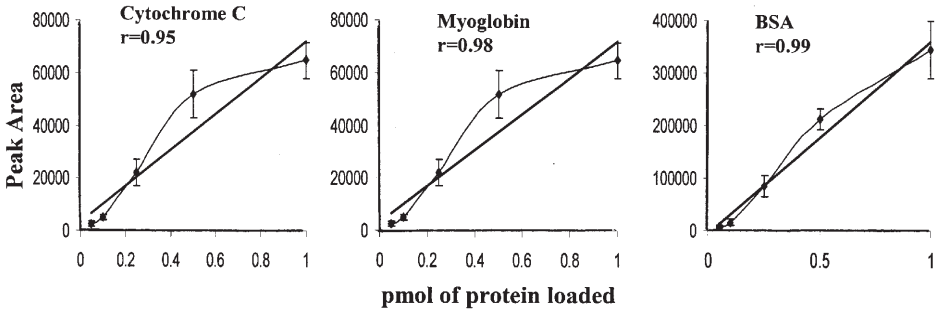

Figure 1. Correlation between amount of pure protein loaded and signal intensity by SELDI-TOF. Three pure proteins, cytochrome C, myoglobin and BSA, at various concentrations were quantified by SELDI-TOF as described in Materials and methods and Results. The correlation between area under the signal and the amount of protein loaded was estimated by Pearson correlation coefficient (r).

Table I. Limits of detection of pure proteins by SELDI-TOF.

Protein	Linear range of detection (pmol)	Correlation coefficient (pmol)	Lowest amount detected (pmol)
Cytochrome C	0.05-1	0.95	0.05
Myoglobin	0.05-1	0.98	0.05
BSA	0.05-1	0.99	0.05

of a 50% (v/v) suspension of the washed beads. Control samples included substitution of 10 μl of distilled/deionized water for the bead suspension. Samples were incubated (4˚C, 20 min) with gentle agitation every 5 min. The supernatant was then removed and 1 μl of each was applied to an NP20 ProteinChip for SELDI-TOF analysis.

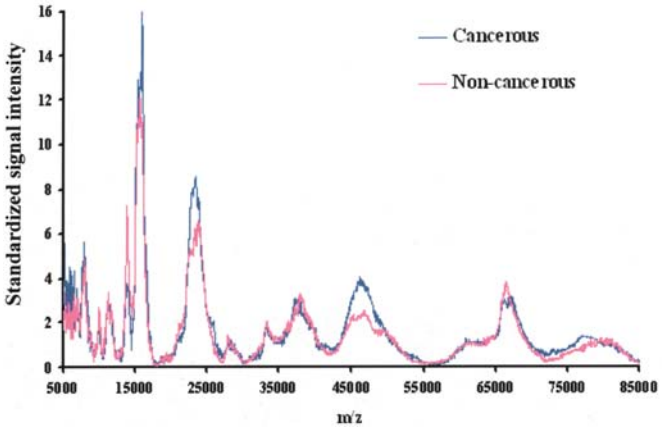

Figure 2. Median intensity SELDI-TOF spectra of NAF from non-cancerous breasts and breasts with invasive cancer. The median intensity at each location of the spectra for NAF from 63 non-cancerous breasts and 38 breasts with invasive cancer is plotted. Pink line, non-cancerous NAF; blue line, NAF from breasts with invasive cancer.

Results

Accuracy of protein quantification by SELDI-TOF. In order to determine the feasibility of using SELDI-TOF analysis for quantitatively comparing the differential target protein abundances, we tested the accuracy of protein quantification by SELDI-TOF using three pure proteins, bovine cytochrome C, myoglobin and serum albumin (BSA) (Michrom Biosources). These proteins were aliquoted to 1 nmol/tube, dried and stored at -80˚C until analysis. The samples were re-dissolved in water to different concentrations, then 0.5 μl was mixed with 0.5 μl freshly prepared sinapinic acid solution (20 mg/ml, 70% acetonitrile/0.1% trifluoroacetic acid) and samples were loaded onto NP20 ProteinChip (Ciphergen Biosystems, Fremont, CA). Allocation of specimens on protein chip arrays was randomized, and in triplicate. Samples on chips were allowed to air-dry at room temperature. The SELDI-TOF analysis was performed, the positive ion laser desorption mass spectra were recorded and then processed by baseline subtraction and normalization as described in Materials and methods. The Pearson correlation coefficient (r) was used to measure the agreement between m/z area under the signal versus the amount of protein loaded (Fig. 1). The data suggests that there is a linear range for pure proteins between 0.05 and 1 pmol. The lowest detection level of pure protein is 0.05 pmol (Table I).

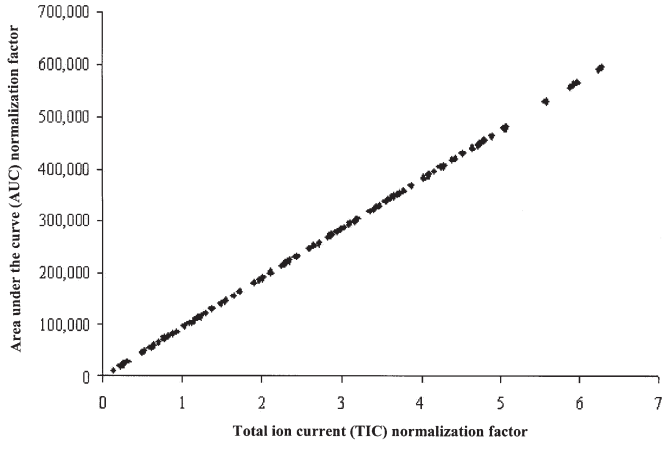

Figure 3. Correlation between TIC and AUC normalizations. Correlation of two normalization methods, TIC and AUC were estimated in the m/z range 5,000-85,000.

SELDI-TOF profiling and statistical analysis. The median SELDI-TOF spectra of NAF from 63 non-cancerous breasts and 38 breasts with invasive cancer revealed a reproducible pattern of peaks in the m/z 5-85 kDa range with apparent differences in relative intensity of signals at several points

Table II. Clusters identified by peak cluster analysis of 101 SELDI spectra.

Cluster no.	m/z	Minimum m/z	Maximum m/z
1	5168[a]	5160	5175
2	6649	6645	6649
3	7582	7574	7593
4	7949	7942	7950
5	10008	9994	10015
6	11350	11345	11359
7	11754	11738	11772
8	12871[a]	12854	12872
9	13543[a]	13526	13559
10	13764[a]	13750	13783
11	13914[a]	13901	13918
12	14033[a]	14022	14049
13	14136[a]	14115	14142
14	15154	15134	15171
15	15334	15314	15342
16	15593	15572	15614
17	15887	15866	15910
18	22630	22602	22661
19	22963	22933	22996
20	23421	23387	23447
21	28024[a]	27983	28065
22	30192	30150	30237
23	30891	30864	30927
24	31722	31694	31758
25	33230	33197	33263
26	38023	37969	38069
27	45714[a]	45661	45754
28	46501[a]	46435	46563
29	61229[a]	61184	61230
30	66727[a]	66652	66813
31	78287	78247	78371
32	79707	79637	79816
33	81309	81197	81422

A cluster was retained when ≥10% of the 101 spectra contributed a peak; [a]potential marker.

(Fig. 2). These differences incited hope that a careful statistical comparison of the two data sets could produce an algorithm that could be useful for cancer diagnosis and/or monitoring the progression of the disease.

The AUC and TIC normalization values were shown to be equivalent (Fig. 3), justifying reporting results for TIC only. This is not surprising since the m/z range is the same for all spectra (m/z 5,000-85,000) and the AUC value approximately equals the mean intensity multiplied by the m/z range, which is the TIC value.

Using normalized spectra, the cluster analysis identified 33 clusters, 12 of which were identified as potential markers (Table II). The median profile analysis involved the examination of 4,851 bins in which 176 potential markers were found (data not shown). All but two of the 12 potential markers identified by the cluster method (61,229, 66,727) were also identified by the median profile method. For each of the 178 potential markers, the accuracy of each one individually ranged from a high of 71% to a low of only 53%. A subset of 14 of the 178 (2+176) potential markers (Table III) had an accuracy of 66% or higher, although the accuracy of the best single potential marker (m/z=5,363) was only 71%. Five of these 14 markers were identified by both median profile and cluster methods. The other nine were identified by median profile only.

The 178 candidate markers were then used as potential predictors in a stepwise logistic regression in order to select a subset of markers that are simultaneously significant and non-redundant. Potential markers were included in the logistic model if their conditional p-value controlling for the other markers was ≤0.15, a more liberal criterion than the $p<0.05$ criterion. The conditional p-value is not the same as the individual p-value. The conditional p-value depends on what other markers are in the model. Eight markers out of 178 were selected by this criterion (Table IV) and were all significant with conditional $p<0.07$.

The composite cancer score based on these eight markers selected by the logistic regression is given by the equation: cancer score = 1.73 + 0.142 M5061 + 0.487 M5994 - 0.491 M6001 - 0.307 M10207 -0.777 M13070 + 2.502 M13436 - 2.714 M13447 - 2.516 M5707, where, for example, M5061 denotes the normalized intensity value at m/z 5061. A positive regression weight implies that, controlling for the other markers in the equation, the conditional marker intensity is higher in the cancer group compared to the non-cancer and thus an excess is associated with an increase in cancer risk. A negative weight implies that the intensity is reduced in the cancer group and is therefore associated with a decrease in cancer risk. While the eight markers chosen by logistic regression in Table IV do not coincide exactly with the 14 markers with highest individual accuracy in Table III, adjacent markers tend to have high correlations with each other and may be partial proxies. For example, while m/z 5,363 is the best individual marker, the correlation of m/z 5,363 with a linear combination of m/z 5,061 and m/z 5,994 from the logistic model is r=0.71. The logistic model is designed to find a set of less redundant (low correlated) markers that can be used simultaneously, not individually, to discriminate cancer from non-cancer even if they are not the best individual discriminators.

The median cancer score based on the above equation in the 38 invasive tumor cancer patients was 0.50 and ranged from -2.5 to 3.9. The median score in the 63 non-cancer patients was -1.16 and ranged from -24.8 to 2.0 ($p<0.001$). While the median is distinctly lower in the non-cancer group, there is still an overlap in their ranges. Using this score to distinguish cancer from non-cancer, the nominal sensitivity was 24/38 (63%) and the nominal specificity was 56/63 (89%) corresponding to an unweighted accuracy of (63%+89%)/2=76%. The ROC curve area was 0.825. While sensitivity decreases when specificity increases for different score thresholds, the ROC statistic is invariant regardless of what score threshold is used.

Table III. Individual markers with an accuracy ≥66% (univariate).

Marker (m/z)	Cancer/non-cancer intensity ratio	Sensitivity (%)	Specificity (%)	Accuracy (%)	p-value
5363	1.75	82	60	71	0.0056
5277	1.79	79	57	68	0.0055
6036	2.33	76	59	68	0.0114
13363	0.73	95	40	67	0.0421
12987	0.52	84	49	67	0.0424
53668	2.08	71	62	66	0.0355
5284	1.7	79	54	66	0.006
5477	1.59	95	38	66	0.0277
5330	1.82	82	51	66	0.0244
5944	2.09	53	79	66	0.0071
13734	0.61	61	71	66	0.0165
13300	0.51	71	60	66	0.0328
5304	1.8	79	52	66	0.0136
5323	1.56	79	52	66	0.0277

Table IV. Simultaneously significant markers from 178 candidate markers (multivariate).

Marker (m/z)	Regression weight	p-value	Individual sensitivity (%)	Individual specificity (%)
5061	0.142	0.0358	63	65
5994	0.487	0.0289	47	83
6001	-0.491	0.0451	53	78
10207	-0.307	0.0420	58	65
13070	-0.777	0.0636	87	41
13436	2.502	0.0145	66	59
13447	-2.714	0.0061	87	41
57075	-2.516	0.0038	95	35

Protein identification. Identification of proteins by SELDI-TOF molecular weight measurements is impossible for at least two reasons. Firstly, the large error in the measurement (±0.2% of the molecular weight) precludes use as an effective screen of lists of protein molecular weights. Secondly, coincidence between calculated molecular weights, derived from amino acid or nucleotide sequences, and measured molecular weights is rare, particularly in eukaryotes because of the frequent occurrence of post-translational modifications (14). Thus reverse-phase chromatography was used to separate NAF proteins. This generated partially purified fractions for protein identification. These fractions were first screened by SELDI-TOF (and MALDI-TOF) so the SELDI-TOF signals from unfractionated NAF could be correlated with fractions from the reverse phase chromatograms. The partially purified samples were treated with a thiol reductant, then an alkylating reagent to block free thiol groups, followed by trypsin digestion, and the resulting mixtures were analyzed by μlC/MS/MS. Proteins were identified by the presence of two or more tryptic peptides with the exception of alipoprotein A-I and A-II that were identified by the presence of a single tryptic peptide. Based on the proteins identified in these experiments, and other experiments where unfractionated NAF was reduced, alkylated and digested prior to analysis by μlC/MS/MS (data not shown), we have compiled a list representing the most abundant proteins detected in NAF (Table V). The intact masses of these proteins were then matched to the potential biomarker m/z values, taking into account the propensity of some proteins to form multiply charged ions as well as non-covalent aggregates (for example protonated dimeric and trimeric ions) during laser desorption. Thus potential biomarkers were given the following tentative identifications (Table V): m/z 14,139, apolipoprotein A-I; m/z 15,091, ß-hemoglobin; m/z 15,268, α-hemoglobin; m/z 15,700, prolactin induced protein; m/z 23,000, apolipoprotein D; m/z 28,143, apolipoprotein A-I; m/z 42,000, Zn α-2-glycoprotein; m/z 50,998, antitrypsin; m/z 53,898, clusterin; m/z 66,664, albumin; m/z 70656, ß-glucuronidase (EC.3.2.1.31) chain A; and m/z 76,016, lactotransferrin.

Protein depletion. A confirmational step was taken to verify that the prominent proteins identified by μLC/MS/MS coincided with the SELDI-TOF peaks seen in the spectra from unfractionated NAF. This was done using an immunoprecipitation method to remove specific proteins from the samples. Following depletion of albumin, NAF was re-analyzed by SELDI-TOF and the resulting spectra were compared to that from the non-depleted control. The m/z peaks at 66,727, 47714, 33,230, and 22630 were no longer detected in the albumin-depleted sample (Fig. 4A). The assignment of all these signals to albumin at different charge states was thus confirmed; the signals at m/z 22629.7, 33229.9 and 47713.9 are attributed to pentuply and doubly charged molecule and a

Table V. Summary of proteins identified in NAF.

Protein name	Swiss-Prot Accession	Comments
Albumin[a]	P02768[c,d]	The main protein of plasma, has a good binding capacity for water, Ca^{2+}, Na^{+}, K^{+}, fatty acids, hormones, bilirubin and drugs. Involved in the regulation of the colloidal osmotic pressure of blood.
Actin, alpha skeletal muscle[b]	P68133[c]	Involved in various types of cell motility and ubiquitously expressed in all eukaryotic cells.
Apolipoprotein A-I[a]	P02647[d]	Major protein of plasma HDL. Participates in the reverse transport of cholesterol from tissues to the liver for excretion.
Apolipoprotein A-II[a]	P02652[c]	May stabilize HDL structure by its association with lipids, and affect the HDL metabolism. Also forms a disulfide-linked heterodimer with apoD.
Apolipoprotein D[a]	P05090[c,d]	Expressed in liver, intestine, pancreas, kidney, placenta, adrenal, spleen, fetal brain tissue and tears. Primarily localized in HDL (60-65%). Involved in the transport and binding of bilin.
ß-Glucuronidase	P08236[c]	Plays an important role in the degradation of dermatan and keratan sulfates.
Clusterin[a]	P10909[c]	Function is not clear. Expressed in a variety of tissues. Associated with apoptosis.
CD59 glycoprotein[a]	P13987[c]	Potent inhibitor of the complement membrane attack complex (MAC). Involved in signal transduction for T-cell activation when complexed to a protein tyrosine kinase. Interacts with T-cell surface antigen CD2.
Complement C4[a]	P01028[c]	Plays a central role in the activation of the classical pathway of the complement system. Circulates in blood as a disulfide-linked trimer of an α, ß and γ chain.
Hemoglobin ß[b]	P68871[d]	Involved in oxygen transport from the lung to the various peripheral tissues.
Hemoglobin α[b]	P69905[d]	Involved in oxygen transport from the lung to the various peripheral tissues.
Ig α-2 chain C region[a]	P01877[c]	Ig α is the major immunoglobulin class in body secretions.
Ig α-1 chain C region[a]	P01876[c]	Ig α is the major immunoglobulin class in body secretions.
Lactotransferrin[a]	P02788[d]	Iron binding transport protein. Has antimicrobial activity that depends on the extracellular cation concentration.
Mucin short variant SV10	Q7Z536[c]	Produced by secretory epithelial cells for ductal and lumenal protection.
Prolactin-induced protein[a]	P12273[c,d]	Expressed in pathological conditions of the mammary gland and in several exocrine tissues, such as the lacrimal, salivary, and sweat glands. Induced by prolactin and androgen; inhibited by estrogen.
Zinc-α2-glycoprotein[a]	P25311[c]	Found in blood plasma, seminal plasma, urine, saliva, sweat, epithelial cells of various human glands, and liver. Stimulates lipid degradation in adipocytes and causes the extensive fat losses associated with some advanced cancers.
α-1-Antitrypsin[a]	P01009[c,d]	Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin.

[a]Known to be secreted and plasmatic; [b]cellular leakage proteins known to be in plasma; [c]protein identified in a benign NAF; [d]protein identified in NAF with invasive tumor present.

triply-charged dimer, respectively. Similarly, hemoglobin depleted and non-depleted NAF samples were prepared and analyzed by SELDI-TOF. The peaks at m/z 15,154 and 15,593 were significantly reduced, and the peaks at m/z 7,582, 7,949, 30,192, 30,891 and 31,720 were absent from the spectrum of the depleted sample (Fig. 4B). Thus the SELDI-TOF peaks at m/z 7,582 and 7,949 were assigned as the doubly charged hemoglobin ions, and the signals at m/z 30,192, 30,891 and 31,720 were assigned as the α/α-dimer, α/ß-heterodimer, and ß/ß-dimer, respectively.

Discussion

A vigorous search for body fluid biomarkers that discriminate between diseased and non-diseased state has been precipitated by the development of analytical methods for comprehensively

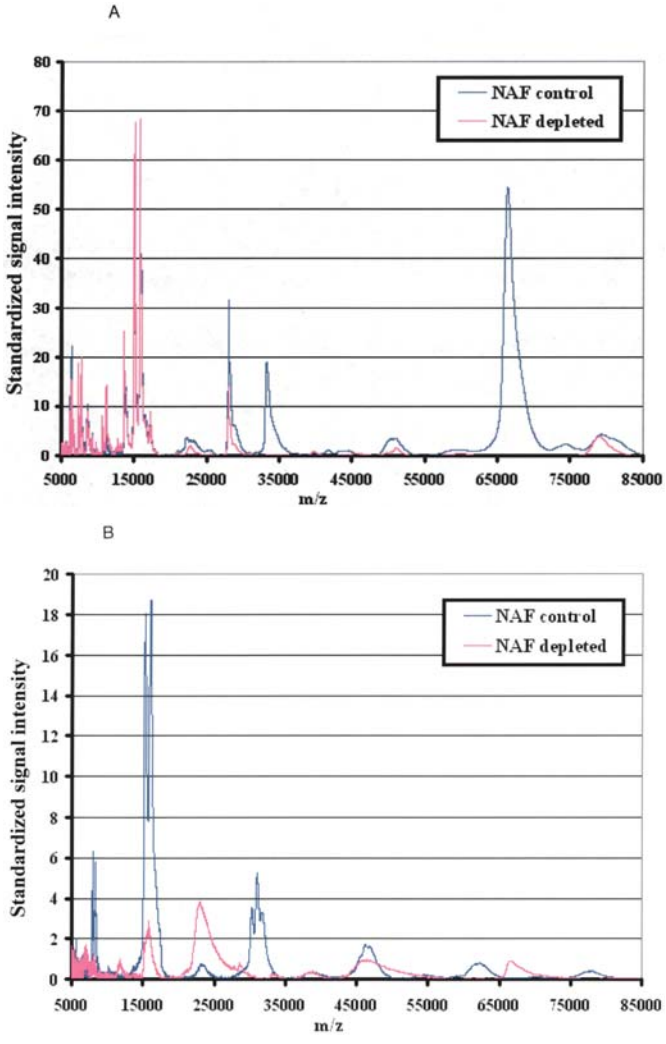

Figure 4. Depletion of abundant proteins from NAF. Depletion of abundant proteins from NAF was performed as described in Materials and methods. A, depletion of albumin; B, depletion of hemoglobin. Pink line, NAF-depleted; blue line, NAF control.

displaying protein and peptide patterns in one and two-dimensional formats. The simplest approach to this search has been taken here with a direct examination of unfractionated NAF by SELDI-TOF to seek markers or a combination of biomarkers that can discriminate between cancerous and non-cancerous breasts. Measurement of three pure proteins by SELDI-TOF analysis suggested a linear quantitative feature, although the detection level and linear range of proteins in serum or other complex specimens is likely to be more complicated. The mass spectra from NAF specimens reveal a complex but reproducible pattern of peaks with qualitative differences between the non-cancer and invasive cancer groups. A special form of statistical analysis has been developed to interpret the results (15) and to produce a statistic (cancer score) based on a linear combination of signals and their intensities that best distinguishes between cancer and non-cancer. The result is promising in that a sensitivity of 63% and specificity of 89% can be assigned to this score at the current state of development of the algorithm. While not yet clinically useful, this procedure appears to hold promise for cancer screening and warrants further investigation and refinement.

NAF is a secretion from the mammary gland that can be collected in a painless, non-invasive out-patient clinical setting. The fluid is derived from the apocrine and merocrine gland-like lobular-ductal system of the breast, which secretes many of the proteins found in milk, including albumin, complement factors and immunoglobulins (16). NAF from cancerous breast may contain secretion unique to cancer cells. Furthermore, NAF may also contain a small number of cancer cells and these cells would be presumably lysed when frozen NAF is thawed. Thus the fluid may contain the cytosolic components of the lysed cells and their membranous fragments. It is reasonable to expect that NAF from a cancerous breast will contain some of the cellular components and secreted proteins of malignant cells in greater relative abundance than these components present in the circulation. In this respect, NAF represents an important source of body fluids for breast cancer biomarker discovery. However, proteomic analysis of NAF has several technical challenges. First, the volume of fluid that can be collected is small and variable (typically 1-10 μl). Second, the protein content ranges widely between 1-90 mg/ml (Bradford protein assay, BSA standard, average 60 mg/ml, data not shown). Lastly, the large variation in fluid viscosity is attributable at least in part to the rich but variable lipid content. Nevertheless, SELDI-TOF spectra of diluted, unfractionated samples can be easily and quickly collected.

However, the processing of multiple SELDI-TOF raw data files, and comparisons between files and groups of files, is an even more complex issue. This involves several inter-related issues, and the currently available software programs for data manipulation proceed through noise estimation, baseline correction and peak finding steps, etc., in a sequence that marginalizes their utility for making global distinctions between two sets of spectra (10). As yet there is no universal generally accepted agreement regarding how to best normalize mass spectra intensity values for the purposes relevant here. Another major concern is whether peak heights either before or after normalization accurately reflect the protein concentration in the applied sample, i.e. whether the SELDI-TOF intensities are quantitative. Improvement in the quantitative nature of SELDI-TOF spectra is an important parameter that would strengthen the utility of this approach and possibly bring the sensitivity and specificity of the cancer screening algorithm to within a clinically useful range.

Both hierarchical and k mean clustering approaches have been previously used for the selection of uniform and non-uniform bins. These different methods yield non-identical results. There is also no apparent consensus on what percentage of the samples must have a peak in a given cluster in order for there to be a useful amount of discrimination. Furthermore, examination of the current data shows that the distribution of intensity values for a given m/z cluster within a given group is skewed. Thus summarization with medians and the use of non-parametric methods are probably more appropriate than summarizing with means and using parametric approaches.

The sensitivities and specificities given in this report are nominal since they have not been validated in another dataset from the same population, or by any statistical validation procedure based on re-sampling. We suspect that our sample size, while appreciable, may still be inadequate. The nominal

accuracy is still only 76% at best showing that there is still substantial misclassification with these 8 markers. Assuming that there actually do exist markers in NAF that distinguish cancer from non-cancer, the modest accuracy reported here may be in part due to lumping persons from different populations into the same group. Ongoing work will incorporate covariate information such as age, tumor size, ER/PR status and treatment exposure into the analyses in attempts to reduce heterogeneity within each group.

Four studies have previously reported on the utility of SELDI-TOF analysis of NAF to distinguish between patients with and without breast cancer (7,9,12,17). Most importantly, Sauter *et al* (9) found signals at m/z 6,500, 15,940, 28,100 and 31,770 Da in a high percentage of samples from breast cancer patients, but in a low percentage of samples from control subjects. Subsequently, protein identification was achieved by an immunoprecipitation approach. This showed that masses 8,000, 15,900, and 31,770 were attributable to ß-hemoglobin. Disappointingly, in this follow-up study, the other signals ear-marked as of-interest in the first study, were not found to be associated with cancer. Pawlik *et al* (17) studied paired NAF samples from 23 women and NAF samples from healthy volunteers. In breast cancer patients, no differences in NAF SELDI signals were identified between the breast with the tumor and the contralateral non-cancerous breast. But differences were observed when spectra from the non-cancerous bearing breast of the breast cancer patients were compared with spectra from breast of healthy volunteers. These studies highlight the difficulties involved in ear-marking potential breast cancer markers and then identifying the responsible proteins.

While the goal of this work is to identify a marker or suite of markers that are useful for disease diagnosis, it is also important that the proteins of interest be identified (7,9,12,17). Off-line chromatographic purification followed by μlC/MS/MS analysis on partially purified fractions was used to match predicted characteristics of entries in the databases with observed MS and MS/MS spectra using the Sequest software package. This resulted in the identification of 17 proteins. Of these, seven have been previously reported to be altered in NAF, serum or tumor tissue from women with breast cancer. Some of the NAF proteins are known blood components. Whether these arise from blood or are true NAF constituents is irrelevant, however, in terms of the goal of the biomarker discovery project (18). Plasma antitrypsin was previously reported to be elevated in subjects with several types of cancer, with a significant increase in breast cancer (19). Serum levels of apolipoprotein A-I have been reported to be significantly higher in women with cancer recurrence than in women without cancer recurrence (20). Apolipoprotein A-II has been associated with breast cancer, but it does not correlate with the recurrence of cancer (20). Apolipoprotein D is present in benign and cancerous NAF, but the level of this protein has been reported to be higher in women without cancer than in women with cancer (21), although a recent study failed to replicate this finding (18). Overexpression of clusterin has been associated with breast carcinoma (22). Zinc-α2-glycoprotein levels in NAF has been reported to be higher in samples from pre-menopausal but not from menopausal women with breast cancer (18). Prolactin induced protein was reported to be lower in NAF from cancerous samples than in samples from breasts without cancer (18). Importantly, a strength of the methods employed here lies in the use of a combination of protein signals to discriminate between samples from cancerous and non-cancerous breasts. This represents a logical progression and evolution of the biomarker discovery strategy beyond the use of single biomarker entities.

It is likely that only abundant to moderately abundant proteins will be detected by SELDI-TOF analysis of un-fractionated NAF (16,18). This problem is shared by all proteomic biomarker screening strategies. The need to access displays of the less abundant sample components emerged with the realization that perhaps the most important information may lie with cell signaling and transcription factors that are present in low intracellular concentrations and in even lower concentrations in body fluids. This need has resulted in the development of antibody-based procedures for removal of the major components. While the application of these procedures poses the risk of simultaneous loss of bound components, their application to NAF analysis is a logical step and will be pursued as this research progresses. In the present study these procedures were used to confirm the identity of some of the major SELDI-TOF signals. Thus some of the strong signals in the SELDI-TOF spectra were shown to be attributable to albumin at various charge and dimerization states and α- and ß-hemoglobins in various oligomerization states.

Acknowledgements

This work was supported by the California Breast Cancer Research Program (6JB-0013), the Department of Defense (DAMD17-01-1-0179), the National Institutes of Health (1RO1CA93736) and the Gonda Foundation to H.R.C.

References

1. Suzuki T, Toi M, Saji S, *et al*: Early breast cancer. Int J Clin Oncol 11: 108-119, 2006.
2. Baines CJ: Are there downsides to mammography screening? Breast J 11 (suppl. 1): S7-S10, 2005.
3. Malatesta M, Mannello F, Bianchi G, *et al*: Biochemical and ultrastructural features of human milk and nipple aspirate fluids. J Clin Lab Anal 14: 330-335, 2000.
4. Krishnamurthy S, Sneige N, Thompson PA, *et al*: Nipple aspirate fluid cytology in breast carcinoma. Cancer 99: 97-104, 2003.
5. Wrensch MR, Petrakis NL, Miike R, *et al*: Breast cancer risk in women with abnormal cytology in nipple aspirates of breast fluid. J Natl Cancer Inst 93: 1791-1798, 2001.
6. Sauter ER, Ehya H, Mammen A, *et al*: Nipple aspirate cytology and pathologic parameters predict residual cancer and nodal involvement after excisional breast biopsy. Br J Cancer 85: 1952-1957, 2001.
7. Paweletz CP, Trock B, Pennanen M, *et al*: Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. Dis Markers 17: 301-307, 2001.
8. Li J, Zhang Z, Rosenzweig J, *et al*: Proteomics and bio-informatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 48: 1296-1304, 2002.
9. Sauter ER, Zhu W, Fan XJ, *et al*: Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. Br J Cancer 86: 1440-1443, 2002.
10. Coombes KR, Fritsche HA Jr, Clarke C, *et al*: Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clin Chem 49: 1615-1623, 2003.
11. Klein P, Glaser E, Grogan L, *et al*: Biomarker assays in nipple aspirate fluid. Breast J 7: 378-387, 2001.

12. Sauter ER, Shan S, Hewett JE, *et al*: Proteomic analysis of nipple aspirate fluid using SELDI-TOF-MS. Int J Cancer 114: 791-796, 2005.
13. Izmirlian G: Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. Ann NY Acad Sci 1020: 154-174, 2004.
14. Strupat K: Molecular weight determination of peptides and proteins by ESI and MALDI. Methods Enzymol 405: 1-36, 2005.
15. Wright GL Jr: SELDI proteinchip MS: a platform for biomarker discovery and cancer diagnosis. Expert Rev Mol Diagn 2: 549-563, 2002.
16. Varnum SM, Covington CC, Woodbury RL, *et al*: Proteomic characterization of nipple aspirate fluid: identification of potential biomarkers of breast cancer. Breast Cancer Res Treat 80: 87-97, 2003.
17. Pawlik TM, Fritsche H, Coombes KR, *et al*: Significant differences in nipple aspirate fluid protein expression between healthy women and those with breast cancer demonstrated by time-of-flight mass spectrometry. Breast Cancer Res Treat 89: 149-157, 2005.
18. Alexander H, Stegner AL, Wagner-Mann C, *et al*: Proteomic analysis to identify breast cancer biomarkers in nipple aspirate fluid. Clin Cancer Res 10: 7500-7510, 2004.
19. Wojtukiewicz MZ, Rucinska M, Kloczko J, *et al*: Profiles of plasma serpins in patients with advanced malignant melanoma, gastric cancer and breast cancer. Haemostasis 28: 7-13, 1998.
20. Lane DM, Boatman KK and McConathy WJ: Serum lipids and apolipoproteins in women with breast masses. Breast Cancer Res Treat 34: 161-169, 1995.
21. Sanchez LM, Diez-Itza I, Vizoso F, *et al*: Cholesterol and apolipoprotein D in gross cystic disease of the breast. Clin Chem 38: 695-698, 1992.
22. Redondo M, Villar E, Torres-Munoz J, *et al*: Overexpression of clusterin in human breast carcinoma. Am J Pathol 157: 393-399, 2000.