SPANDIDOS PUBLICATIONS

Identification of a 17-protein signature in the serum of lung cancer patients

ROMAN T. SRESELI[1*], HARALD BINDER[3*], MADELEINE KUHN[1], WERNER DIGEL[1] HENDRIK VEELKEN[1], WULF SIENEL[2], BERNWARD PASSLICK[2], MARTIN SCHUMACHER[3], UWE M. MARTENS[1**] and STEFAN ZIMMERMANN[1**]

Departments of [1]Hematology and Oncology, and [2]Thoracic Surgery, University Medical Center Freiburg, Hugstetter Strasse 55, D-79106 Freiburg; [3]Department of Medical Biometry and Statistics, University Medical Center Freiburg, Stefan Meier Strasse 26, D-79104 Freiburg, Germany

Received January 26, 2010; Accepted March 18, 2010

DOI: 10.3892/or_00000855

Abstract. Early detection of lung cancer may potentially help to improve the outcome of this fatal disease. Currently, no satisfactory laboratory tests are available to screen for this type of cancer. The aim of this study was to improve diagnostic procedures for lung cancer through the discovery of serum biomarkers using SELDI-TOF MS (surface-enhanced laser desorption/ionization time-of-flight mass spectrometry). Mass spectrometric profiling was applied to the serum of a total of 139 lung cancer patients and 158 healthy individuals for developing a prognostic signature. For validation, two separate groups were employed, comprising of 126 and 50 individuals, respectively. Informative regions of mass spectra were identified by forming protein mass clusters and identifying predictive clusters in a logistic regression model. A total of 17 differential predictive protein mass clusters were identified in patients with metastatic lung cancer disease compared to healthy individuals. These clusters provide for a robust risk prediction model. The sensitivity and specificity of this model was estimated to be 87.3 and 81.9%, respectively, for the first validation set, and 96.0 and 66.7%, respectively, for a second validation set of patients with early disease (stages I and II). A differential 11.5/11.7 kDa double-peak could be identified as serum amyloid α (SAA) by peptide mapping. Our data suggest that the SELDI-TOF MS technology in combination with a careful statistical analysis appears to be a useful experimental platform which delivers a rapid insight into the proteome of body fluids and may guide to identify novel biomarkers for lung cancer disease.

Correspondence to: Dr Stefan Zimmermann, Department of Hematology and Oncology, University Medical Center Freiburg, Hugstetter Strasse 55, D-79106 Freiburg, Germany
E-mail: stefan.zimmermann@uniklinik-freiburg.de

[*]Contributed equally

[**]Shared senior authorship

Key words: lung cancer, proteomics, biomarker, SELDI-TOF MS

Introduction

With a high incidence and a total 5-year survival rate of only 10-15%, lung cancer (LC) remains the most common cause of cancer-related death in Europe and in the USA (1). For both major LC subtypes, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), the probability of cure is predominantly a function of the extent of the anatomical spread or clinical stage of the tumor. In NSCLC, which accounts for approximately 80% of LC patients (2), complete surgical resection remains the mainstay of curative therapy but is restricted to patients with stage I-IIIa disease (3). Both early LC detection and accurate recognition of tumor dissemination are important means to increase cure rates and to avoid unnecessary overtreatment of LC patients.

At present, there are no laboratory tests that permit reliable early detection of LC or indicate LC dissemination. Currently used serum tumor markers [i.e. neuron-specific enolase (NSE) (4), carcinoembryonic antigen (CEA), cytokeratin 19 fragments (CYFRA 21-1), and others], as single biomarkers or in combination, lack sufficient sensitivity and/or specificity to fulfil the requirements for population screening and early diagnosis of asymptomatic LC patients (5-9).

Proteomic technologies are used for global profiling and identification of disease-associated markers in biological fluids (10-12). The study of protein expression patterns by mass spectrometry (MS) in tissue or serum offers the opportunity to unravel and understand the molecular complexity of human tumors and has the potential for improvements in resolution, detectability and diagnostic accuracy (13,14). The surface-enhanced laser desorption and ionization technology allows for rapid identification of proteomic patterns in tissues and body fluids (11) by on-chip protein fractionation coupled to time-of-flight separation by mass spectrometry (SELDI-TOF MS). Several studies

Table I. Clinico-pathological characteristics of patients with lung cancer and healthy controls in training set, validation set I (advanced lung cancer), and validation set II (early-stage lung cancer).

Set	Group	Total (n)	Histology (n)	Stage	Male/Female	Age range (Median)
Training set			Adenocarcinoma (54)			
			Squamous cell lung cancer (33)			
	Lung cancer	139	Small cell lung cancer (31)	III-IV	101/38	35-86 (64)
			Large cell lung cancer (11)			
			Undifferentiated lung cancer (10)			
	Control	158	-	-	93/65	21-69 (37)
Validation set I			Adenocarcinoma (34)			
			Squamous cell lung cancer (16)			
	Lung cancer	63	Small cell lung cancer (7)	III-IV	43/20	37-86 (68)
			Large cell lung cancer (4)			
			Undifferentiated lung cancer (2)			
	Control	63	-	-	37/26	20-68 (35)
Validation set II			Adenocarcinoma (15)			
	Lung cancer	25	Squamous cell lung cancer (9)	I-II	23/2	52-81 (68)
			Large cell lung cancer (1)			
	Control	25	-	-	16/9	21-67 (39)

have suggested the usefulness of SELDI-TOF MS for the identification of novel cancer-associated biomarkers that may permit diagnosis in early disease stages and monitoring for tumor recurrence (4,15-20).

The aim of this study was to investigate proteomic expression differences between lung cancer patients and a healthy population by SELDI-TOF MS technology in order to discover novel LC-associated biomarkers.

Materials and methods

Patients and controls. Serum samples from LC patients at initial diagnosis and healthy volunteers were processed according to a standardized protocol within 1.5 h after phlebotomy at room temperature. Venous blood (5 ml) was centrifuged at 3500 rpm for 10 min, distributed into 200 μl aliquots, and stored at -80˚C.

The training set included 139 lung cancer patients and 158 healthy individuals (Table I). A validation sample set (validation set I) was collected from 63 lung cancer patients and 63 controls. All patients of the training and validation set I had advanced disease stage (UICC stages III and IV). In addition, 25 samples from patients with early stage lung cancer (UICC stages I and II) and additional 25 control samples from healthy individuals were collected (validation set II). All probands gave informed consent, and the study was approved by the institutional ethics committee.

Sample preparation and SELDI-TOF MS. Each (9 μl) serum sample was mixed with 13.5 μl of urea buffer (9 mM urea, 2% CHAPS, 50 mM TRIS-Cl, pH 9.0), placed on an air shaker for 20 min in ice for proper denaturation, after which 200 μl of binding buffer (100 mM sodium acetate, pH 4.0) was added. Initial analyses included CM10 weak cation exchange, Q10 strong anion exchange, and Cu-IMAC (copper-immobilized metal affinity chromatography) Protein Chips (Ciphergen Biosystems, Inc., Fremont, CA). Protein Chips were processed in a Bioprocessor by adding 150 μl binding buffer to each well, incubating for 5 min at RT with shaking at 700 rpm. After repetition of this step, the buffer was removed, and 100 μl denatured protein sample was added immediately in duplicate. After 30 min on a platform shaker at RT, the arrays were washed 3 times with 150 μl binding buffer for 5 min followed by two quick rinses with 200 μl of de-ionized water. The arrays were dried for 15-20 min. Finally, 1 μl of half-saturated SPA dissolved in 50% ACN/ 0.5% TFA was added twice. ProteinChips were stored at RT in the dark until analysis by a ProteinChip reader (Protein Biology System IIc, Ciphergen Biosystems) in two runs in a low mass (LM) and a high mass (HM) range. Laser intensity was set at 170 for LM (195 for HM) and detector sensitivity at 7 (9 for HM) to acquire an optimal mass from 2 to 30 kDa (10-50 kDa) and a maximum mass of 50 kDa (150 kDa). Spectra consisted of 130 averaged laser shots. External calibration was performed using an All-in-One Protein Standard II (hirudin BHVK, 7.0 kDa; bovine cytochrome C, 12.2 kDa; equine myoglobin, 17.0 kDa; bovine carbonic anhydrase, 29.0 kDa; yeast enolase, 46.7 kDa; bovine albumin, 66.4 kDa; and bovine IgG, 147.3 kDa; Ciphergen Biosystems). Automatic baseline correction was used and peaks with an m/z value of 2.5 kDa were mainly ion noise from the matrix and were therefore excluded from the analysis. The peak intensities were normalized to the total ion current of m/z between 2.5 and 150 kDa, and data filtering was set as default at 0.2 times of the expected peak width. The data were pre-processed with ProteinChip Software 3.1

n Biosystems, Inc.). Further protein peak clustering sification analyses were made using Ciphergen Express- and Biomarker Pattern software (all Ciphergen Biosystems).

Statistical analysis. To extract covariates for subsequent analyses from individual protein spectra of the training set, peaks were defined through the package 'ppc', employing a minimal signal-to-noise ratio 3, peak span 0.05 on log scale, and smoothing span 0.5 on log scale (21), in the statistical environment R (22). A predictive model was then constructed by means of a logistic regression model with penalized parameter estimation (23). Analysis decisions that were found to be required in the course of model building were based on prediction performance estimated from 10-fold cross-validation. This approach permitted to build the model exclusively on the training set and saved the validation set for controlling the final model. This principle of analysis is important because identification of peak clusters from the combined data would have biased the estimated prediction error of a later final model on the validation set in downward direction. Also, modelling of interactions, as provided e.g. by decisions tree modelling techniques or support vector machines, was deliberately not considered to allow clear interpretations.

In the first step of model building, hierarchical clustering of defined peaks was performed based on a maximal cluster width of 0.1 on log scale (21). Covariate values of 0 or 1 were determined for each cluster and for each spectrum according to the presence or absence of a peak in the respective cluster within a given spectrum. The actual peak intensities were deliberately not used since a large variability of intensities is known to occur in repeated measurements, and cross-validation indicated that the predictive performance was influenced negligibly by this decision.

Based on the defined covariates (which correspond to the peak clusters) and a 0/1 response that indicated whether a sample was from a cancer patient (value 1) or from a healthy control (value 0), the parameters of a logistic regression model were estimated. For each of the peak clusters, one estimated parameter was obtained that indicates whether the respective peak cluster was deemed informative in a multivariate model, i.e. whether the information of a sample having a peak in the respective cluster was useful for predicting the class membership (cancer vs. healthy) of the sample. Non-informative clusters were identified by close-to-zero or ideally equal to zero parameter estimates.

To reduce variability of estimates and to increase model stability, penalized estimation was applied. We chose a Lasso-like L1 penalty on the parameters (23-25) because this leads to sparse models, i.e. only a subset of the peak clusters receives non-zero parameter estimates while maintaining good predictive performance. This procedure has one tuning parameter, the amount of penalization, that in effect determines what number of peak clusters receive non-zero parameter estimates, i.e. non-zero influence. An advantage of this specific procedure [path algorithm for generalized linear models (23), implemented in the R package software 'glmpath'] is the availability of information criteria for automatic selection of this parameter, whereas computationally more intensive and potentially more unstable resampling methods would have been necessary otherwise. We employed the Bayesian information criterion (BIC) which is expected to result in conservative models with relatively few influential peak clusters. To note, one disadvantage of the chosen estimation technique is that no standard error estimates, i.e. also no p-values, are obtained for the parameter estimates. Therefore, it is difficult to judge for peak clusters that receive non-zero parameter estimates how important they are. We attempted to obtain a rough approximation by fitting an unpenalized logistic regression model including only those peak clusters that receive non-zero parameter estimates from penalized estimation.

Protein purification, passive elution, and peptide mapping. Proteins of interest were enriched by anion exchange chromatography using Q HyperD F Spin Columns (Ciphergen Biosystems) according to the manufacturer's instructions. Briefly, denatured serum samples were diluted 1:1 in 50 mM Tris-HCl, pH 9.0 and incubated with the equilibrated anion exchange sorbent of the column on a shaker at room temperature for 30 min. Proteins were eluted in 100 μl aliquots by a pH gradient in five decrement steps (pH 9.0, 7.0, 5.0, 4.0, and 3.0) concluded by an organic washing step in a solution of 33.3% isopropanol, 16.7% ACN, and 0.1% TFA. Each fraction (10 μl) diluted in 90 μl of binding buffer was monitored by SELDI-TOF MS on CM10 arrays as described above. The residual of each fraction was separated on 10-16% Tris-tricine polyacrylamide gels that were stained overnight with colloidal Coomassie brilliant blue G250 and destained with 20% methanol. Proteins of low molecular weight (<20 kDa) were passively eluted from gel bands using a solution of 45% formic acid, 30% ACN, 10% isopropanol, 15% H_2O. A small portion (3 μl) of eluted protein was re-analysed by SELDI-TOF MS on normal phase (NP20) arrays. The remaining eluate was lyophilized and resolved in 5 μl of 20 ng/μl sequencing grade trypsin (Promega, Madison, WI) in 25 mM ammonium bicarbonate and 10% ACN and incubated overnight at 37˚C. Tryptic peptides were analysed by SELDI-TOF MS on NP20 arrays using an CHCA matrix and external calibration by an All-in-One Peptide Standard (vasopressin, 1.08 kDa; somatostatin, 1.64 kDa; dynorphin, 2.15 kDa; ACTH 1-24, 2.93 kDa; bovine insulin ß-chain, 3.50 kDa; human insulin, 5.81 kDa; hirudin BHVK, 6.96 kDa; Ciphergen Biosystems). The MASCOT software (Matrix-Science, London, UK) was used in the Peptide Mass Fingerprint mode in combination with the NCBInr database (National Center for Biotechnology Information, WA) for protein identification (search criteria: monoisotopic m/z, mass accuracy 1.2 Da or better, up to one missed cleavage). Probability based Mowse scores >65 were considered as significant ($p<0.05$).

Results

Analysis of mass spectrometry data. Serum samples of untreated lung cancer patients and healthy individuals were analyzed on ProteinChip arrays using SELDI-TOF MS. Prominent differences in the expression of protein peaks

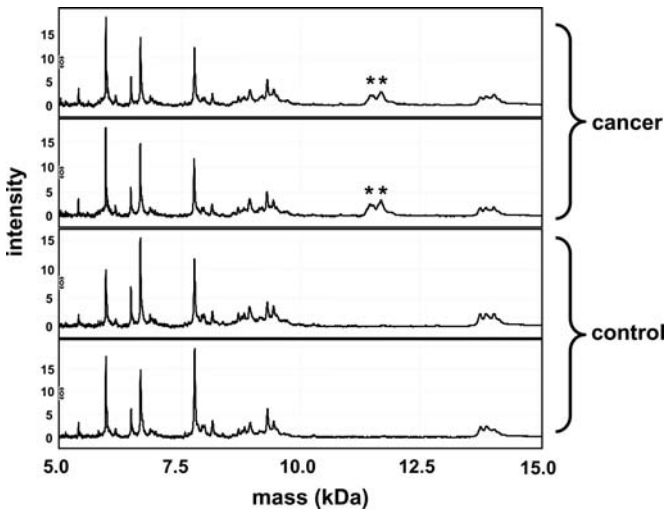

Figure 1. SELDI-TOF MS profiles of representative serum samples in duplicates generated on CM10 arrays. Asterisks indicate peaks at 11.5 and 11.7 kDa evident in the lung cancer samples but not in the controls.

were observed between the two sample groups (Fig. 1). Statistical analysis of protein spectra obtained with the CM10 weak cation exchange surface resulted in 24 peak clusters. Differences in the serum proteomes between normal individuals and advanced lung cancer patients were analyzed by L1-penalized logistic regression with penalty parameter chosen by BIC.

Seventeen of the 24 CM10 peak clusters were identified to distinguish between cancer patients and healthy individuals (Table II). The standard error estimates and p-values obtained from unpenalized estimation indicate the relative importance of the peak clusters. To note, these standard errors and p-values have to be interpreted with caution, not only because they exhibit potential bias because of estimation after peak cluster selection, but also because the unpenalized estimate was rather unstable, e.g. as indicated by some unrealistically large estimates.

The prediction for new samples was based on the very stable penalized estimates. For example, for a sample that exhibits peaks in clusters 4.213, 6.634, 7.768, 8.924, 9.292, and 132.860 Da (but not in the others) the predicted probability of being a cancer sample would be $1/\{1+\exp[-(-1.941+0.415+1.767+0.920+(-1.130)+(-0.544)+3.095)]\} = 1/[1+\exp(-2.582)] = 0.930$ (using the estimated intercept value of -1.941). With a cut-off of 0.5, this sample would be classified correctly to be from a cancer patient.

Since the 17 informative peaks were identified to be influential by a multivariate technique, not every single peak is expected to expose large differences between groups (e.g. the peak at 132.860 Da). For comparison, a peak list based on univariate techniques was created with Ciphergen Express software comparing all different clusters of a given range of profiles, operating p-value calculations

Table II. Peak clusters used in the final predictive logistic regression model.

Peak cluster			Peak performance			Peak prevalence (%)						
Position (Da)	Minimum (Da)	Maximum (Da)	Coefficient	Standard error	p-value	All patients	Adeno-carcinoma	SCC	LCLC	Poorly differentiated	SCLC	Healthy
2.856	2.728	2.867	-1.185	0.985	0.043	19.1	27.4	15.6	22.7	22.2	22.7	24.8
2.946	2.940	3.026	-0.826	0.937	0.434	12.1	17.9	9.4	13.6	22.2	13.6	24.8
3.156	3.076	3.262	-2.784	0.459	<0.001	40.1	36.8	37.5	40.9	38.9	40.9	62.5
3.366	3.306	3.441	3.035	0.611	<0.001	27.2	23.6	25.0	31.8	38.9	31.8	1.6
3.881	3.874	3.974	1.209	1.131	0.216	8.5	3.8	15.6	0	0	0	0.6
4.213	4.197	4.360	0.415	0.463	0.040	66.5	79.2	62.5	59.1	50	59.1	78.7
4.640	4.466	4.647	0.789	0.969	0.138	12.5	3.8	23.4	9.1	5.6	9.1	2.2
5.907	5.895	5.921	1.411	1.175	0.001	56.6	51.9	71.9	54.5	72.2	54.5	41.9
6.634	6.621	6.647	1.767	1.159	<0.001	42.3	47.2	26.6	45.5	27.8	45.5	50.2
7.768	7.753	8.220	0.920	0.786	0.655	94.9	93.4	100	86.4	100	86.4	90.5
8.924	8.705	8.951	-1.130	0.396	0.001	83.1	75.5	95.3	72.7	83.3	72.7	88.3
9.292	9.184	9.440	-0.544	1.263	0.174	97.4	98.1	100	86.4	100	86.4	98.1
11.635	11.476	11.673	4.269	886.0	0.983	35.3	21.7	43.8	45.5	66.7	45.5	0
15.080	14.637	15.830	0.673	1.117	0.363	2.6	1.9	3.1	4.5	0	4.5	0.6
17.317	17.187	17.359	-0.002	0.302	0.756	22.1	41.5	9.4	13.6	11.1	13.6	43.5
28.023	27.966	28.090	-2.208	0.343	<0.001	31.6	35.8	23.4	54.5	11.1	54.5	84.4
132.860	132.330	133.646	3.095	10754.0	0.998	100	100	100	100	100	100	99.7

Coefficients are derived from penalized estimation. Approximate standard errors and p-values are derived from unpenalized estimation after peak selection. SCC, squamous cell cancer; LCLC, large cell lung cancer; SCLC, small cell lung cancer.

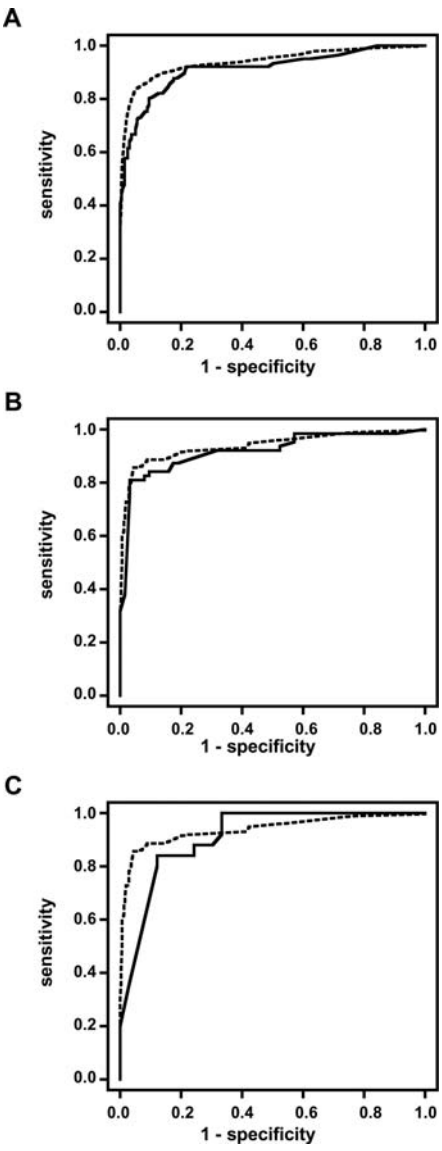

Figure 2. ROC curves showing sensitivities and specificities for varying cut-offs on predicted probabilities. (A) Estimated from 10-fold cross-validation (mean performance on training subsets in cross-validation iterations, dashed curve; mean performance on test subsets in cross-validation iterations, solid curve). (B) For the full training (dashed curve) and the validation set (solid curve). (C) For the full training set (dashed curve) and the set with early stage lung cancer samples (solid curve).

and graphical presentations. Only one additional potentially relevant peak at 13.75 kDa not pointed out by the multivariate analysis was identified (data not shown).

Class prediction performance. ROC curves were generated that indicated sensitivities and specificities for class prediction by smoothly varying the cut-off applied to the predicted probabilities for classification (Fig. 2A). One decision based on cross-validation results was to restrict the analysis to the CM10 data, because addition of the Q10 and Cu-IMAC data did not increase predictive performance much, but considerably increased the number of peaks used by the model for prediction (data not shown).

The comparison to the mean performance on the cross-validation training sets shows that prediction performance is only slightly overestimated in the training data, indicating that the procedure used harbours little danger of overfitting.

The ROC curves for the full training and validation set I were found to be very similar to the results from cross-validation (Fig. 2B; training set AUC: 0.938, validation set AUC: 0.918). This finding indicates that the statistical model employed provides a good prediction performance. Furthermore, it indicates that the structure found in the SELDI-TOF MS data is sufficiently stable to be applied samples obtained at a later date.

The rate of correct classification with a cut-off of 0.5 (applied to the predicted probability) in validation set I was 0.841 (training set: 0.910). The corresponding sensitivity and specificity were 87.3 and 81.9%, respectively. The 15.9% misclassified samples comprise 12 of the 63 healthy controls, 1 of 28 adenocarcinoma patients, 4 of 21 squamous cell cancer patients, 2 of 7 small-cell lung cancer patients, and 1 of 3 undifferentiated or poorly differentiated lung cancer patients. None of the 4 large-cell lung cancer patients was misclassified.

In order to further examine the discriminatory power of the identified peak cluster in lung cancer patients we analyzed the sera of patients with early disease (stages I and II). While the estimate of prediction performance is limited by the small number of samples, a sensitivity of 96.0% and a specificity of 66.7% for a cut-off of 0.5 could be obtained nevertheless (Fig. 2C), which is also reflected in the correct classification rate of 0.793.

Peak identification. For the purification and identification of differentially expressed proteins, serum samples were used which demonstrated the relevant peaks in high abundance. Such samples were fractionated and separated on TRIS-tricine polyacrylamide gels. Proteins of low molecular weight <20 kDa could be passively eluted from gel bands and were re-analysed on an NP20 array. The re-analysis of an eluted protein confirmed its mass identity with the protein peak of the original SELDI-TOF MS spectrum (Fig. 3). In accordance with the original double peak found on Q10 arrays of control samples, a double-peak of 11.5/11.7 kDa was obtained. A Mascot database search for peptide masses from an NP20 array after trypsin digestion of this protein revealed serum amyloid α (SAA; gi:225986), with a significant probability based mowse score of 78 and a sequence coverage of 61%. A molecular weight of 11.675 Da for this isoform of SAA supports the observed peak mass in the SELDI-TOF MS analysis.

Discussion

Mass-spectrometry-driven global proteomic analyses are considered to be key developments for the rapid detection of cancer-specific biomarkers (11,26). In this study, we have identified a serum proteomic classifier for the detection of lung cancer based on a novel sophisticated bioinformatic approach. Clustering of peaks in combination with penalized estimation of a multivariable classification model resulted in a series of biomarkers with a good prediction performance as demonstrated in two independent validation cohorts. As illustrated, the resulting logistic regression model is easily

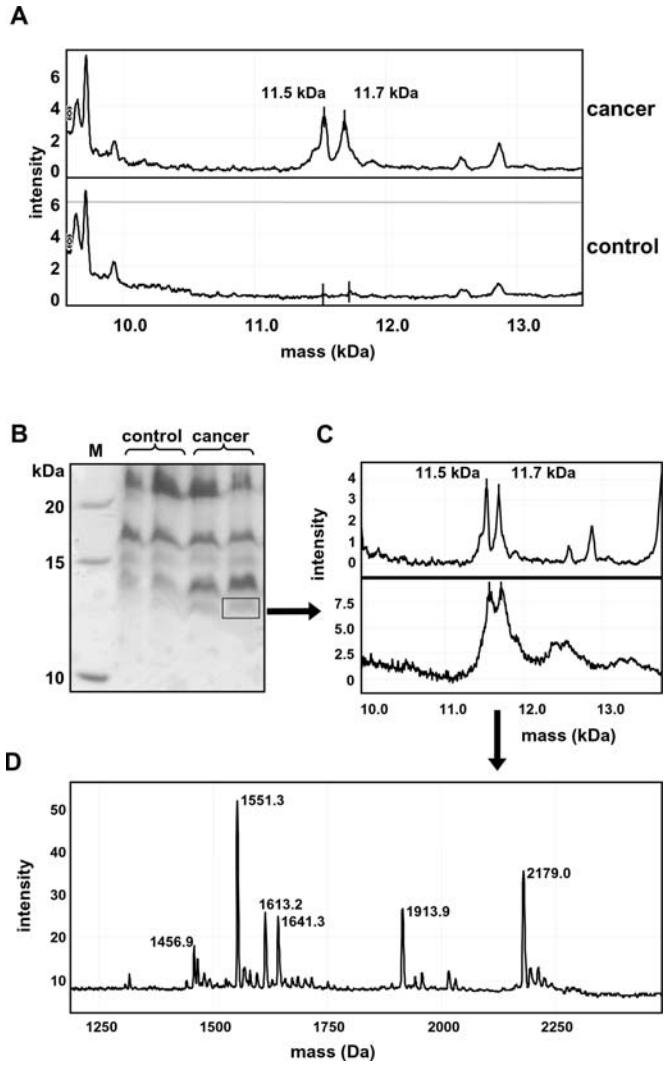

Figure 3. (A) Differential expression of 11.5/11.7 kDa double-peak between representative lung cancer and normal serum (control) samples on CM10 arrays. (B) A band about 12 kDa (organic fraction in TRIS-tricine polyacrylamide gel) was excised and subjected to passive elution and analysed on a normal phase (NP20) array. (C) Passive eluate from 12 kDa band (lower spectrum) could be correlated with 11.5/11.7 kDa double-peak from original CM10 arrays (upper spectrum). (D) NP20 array analysis of peptides of the passive eluted protein after trypsin digestion reveals fingerprint characteristic for Serum Amyloid Alpha (SAA). M, protein marker.

interpretable and returns predicted probabilities of class membership (patient vs. healthy subject). Given the estimated parameters, predictions for new samples were readily obtained by inserting their spectrum/peak information. Importantly, the model discriminated not only patients with advanced lung cancer disease but also patients with early cancer (stages I and II). Proteomic profiles may therefore permit to develop highly sensitive diagnostic tools for the early detection of cancer.

Membership classification, i.e. discrimination of patients with lung cancer from healthy controls, was based on a 17-peak signature (2.856, 2.946, 3.156, 3.366, 3.881, 4.213, 4.640, 5.907, 6.634, 7.768, 8.924, 9.292, 11.635, 15.080, 17.317, 28.023, 132.860 Da). Based on p-value statistics, 6 most relevant differential proteins could be identified (at 3.156, 3.366, 5.907, 6.634, 8.924 and 28.023 Da), however, the model is based on the performance of all 17 proteins presented in the peak list.

Besides the potential for good prediction performance, the multivariate model was expected to identify selectively influential peaks that could not be detected by univariate techniques. For example, compared to the healthy controls, the peak cluster at 6.634 Da is present in a larger proportion for the small-cell lung cancer patients, but in a much smaller proportion for other subtypes. Only a multivariate model can combine this with the information on presence of other peaks to extract the discriminative information from this peak cluster.

The robustness of the interpretation of the proteomic signature is based on the reliability of SELDI-TOF MS technology in combination with the bioinformatics approach. Specifically, prediction for a new patient relies only on the qualitative information of presence/absence of peaks and does not require the quantitative peak height. This feature of the analysis is expected to provide a general level of robustness also for additional applications settings and may help to counter the general scepticism about the reliability of MS-based serum proteomics as cancer biomarkers (27,28).

Several studies have been performed concerning proteomic fingerprinting in cancerous and non-cancerous diseases with identification of specific new biomarkers. Interestingly, some of the differentially expressed proteins described here have been also found by others: 11.6 kDa (29-32), 3.2 and 3.3 kDa (33), 15.1 kDa (34,35), and 17.2 kDa (34), although there are various different proteins patterns identified by each group. Particularly, the 11.6 kDa mass has been identified as serum amyloid A (SAA) (33), which is a major acute phase protein that is associated with circulating high-density lipoprotein and appears to be a potential useful biomarker to monitor relapse of nasopharyngeal cancer. SAA is also known to act as a cytokine and to influence cell adhesion, migration, proliferation, and aggregation. This marker was elevated in 41% of lung cancer patients at initial diagnosis. Furthermore, this protein is one of 20 peaks in another classifier identified through proteomic fingerprinting by SELDI-TOF MS in sera from patients with tuberculosis (36). SAA has recently been identified to represent a prognostic marker in melanoma (37).

Apart from proteomic fingerprinting in sera or plasma from lung cancer patients proteomic patterns have been directly obtained from small amounts of fresh frozen lung-tumour tissue. High-throughput technologies have the potential to view signal transduction networks more globally than it is possible with immunohistochemical analysis (38) and to define distinct prognostic signatures (39,40). However, proteomic profiling from blood allows repeated measurements even during treatment without the need for tumor tissue. Interestingly, a classification algorithm based on MALDI MS analysis was recently reported in sera or plasma that could identify subgroups of NSCLC patients with improved time to progression and overall survival after treatment with the EGFR TKIs gefitinib and erlotinib (41).

In general, detection of lung cancer at early stage disease is an important goal for molecular diagnostics. The predictive performance in our study appears to be quite reasonable with

ity of 87.3% and a specificity of 81.9%. In two ent reports, Yildiz *et al* (32) describe a sensitivity of 58% and specificity of 85.7%, and Jacot *et al* (42) found even a higher performance with a sensitivity of 94.3% and a specificity of 85.9%. Ideally, a simple blood test using only a few biomarkers may represent the most convenient way to improve early detection of lung cancer as described recently (43) by analyzing four proteins: carcino-ambryonic antigen, retinol binding protein, α1-antitrypsin, and squamous cell carcinoma antigen. These markers classified lung cancer patients with 77.8% sensitivity and 75.4% specificity. Nevertheless, it will probably take a combination of novel imaging methods and biomarkers to improve early diagnosis of high-risk individuals with early lung cancer.

In summary, we conclude that serum protein profiling using SELDI-TOF MS combined with a sophisticated bioinformatic approach is a promising tool for diagnosing and early detection of lung cancer patients. Because lung cancer is a highly heterogeneous disease, further studies are necessary to validate the classifier in different histological subsets, smokers and non-smokers, and other non-cancerous lung diseases.

Acknowledgments

We gratefully acknowledge the excellent technical assistance from Sangeeth Sundararajan.

References

1. Jemal A, Siegel R, Ward E, Murray T, Xu J and Thun MJ: Cancer Statistics, 2007. CA Cancer J Clin 57: 43-66, 2007.
2. Wahbah M, Boroumand N, Castro C, El-Zeky F and Eltorky M: Changing trends in the distribution of the histologic types of lung cancer: a review of 4,439 cases. Ann Diagn Pathol 11: 89-96, 2007.
3. Winton T, Livingston R, Johnson D, *et al*: Vinorelbine plus cisplatin vs. observation in resected non-small-cell lung cancer. N Engl J Med 352: 2589-2597, 2005.
4. Liu AY, Zhang H, Sorensen CM and Diamond DL: Analysis of prostate cancer by proteomics using tissue specimens. J Urol 173: 73-78, 2005.
5. European Group on Tumour Markers: Tumour markers in lung cancer: EGTM recommendations. Anticancer Res 19: 2817-2819, 1999.
6. Hirsch FR, Franklin WA, Gazdar AF and Bunn PA Jr: Early detection of lung cancer: clinical perspectives of recent advances in biology and radiology. Clin Cancer Res 7: 5-22, 2001.
7. Kulpa J, Wojcik E, Reinfuss M and Kolodziejski L: Carcino-embryonic antigen, squamous cell carcinoma antigen, CYFRA 21-1, and neuron-specific enolase in squamous cell lung cancer patients. Clin Chem 48: 1931-1937, 2002.
8. Etzioni R, Urban N, Ramsey S, *et al*: The case for early detection. Nat Rev Cancer 3: 243-252, 2003.
9. Basil CF, Zhao Y, Zavaglia K, *et al*: Common cancer biomarkers. Cancer Res 66: 2953-2961, 2006.
10. Pandey A and Mann M: Proteomics to study genes and genomes. Nature 405: 837-846, 2000.
11. Wulfkuhle JD, Liotta LA and Petricoin EF: Proteomic applications for the early detection of cancer. Nat Rev Cancer 3: 267-275, 2003.
12. Ping P, Vondriska TM, Creighton CJ, *et al*: A functional annotation of subproteomes in human plasma. Proteomics 5: 3506-3519, 2005.
13. Diamandis EP: Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics 3: 367-378, 2004.
14. Meyerson M and Carbone D: Genomic and proteomic profiling of lung cancers: lung cancer classification in the age of targeted therapy. J Clin Oncol 23: 3219-3226, 2005.
15. Wilson LL, Tran L, Morton DL and Hoon DS: Detection of differentially expressed proteins in early-stage melanoma patients using SELDI-TOF mass spectrometry. Ann N Y Acad Sci 1022: 317-322, 2004.
16. Kozak KR, Su F, Whitelegge JP, Faull K, Reddy S and Farias-Eisner R: Characterization of serum biomarkers for detection of early stage ovarian cancer. Proteomics 5: 4589-4596, 2005.
17. Liu W, Guan M, Wu D, Zhang Y, Wu Z, Xu M and Lu Y: Using tree analysis pattern and SELDI-TOF-MS to discriminate transitional cell carcinoma of the bladder cancer from non-cancer patients. Eur Urol 47: 456-462, 2005.
18. Ricolleau G, Charbonnel C, Lode L, *et al*: Surface-enhanced laser desorption/ionization time of flight mass spectrometry protein profiling identifies ubiquitin and ferritin light chain as prognostic biomarkers in node-negative breast cancer tumors. Proteomics 6: 1963-1975, 2006.
19. Zhu LR, Zhang WY, Yu L, Zheng YH, Zhang JZ and Liao QP: Serum proteomic features for detection of endometrial cancer. Int J Gynecol Cancer 16: 1374-1378, 2006.
20. Engwegen JY, Helgason HH, Cats A, Harris N, Bonfrer JM, Schellens JH and Beijnen JH: Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry. World J Gastroenterol 12: 1536-1544, 2006.
21. Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A and Le QT: Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics 20: 3034-3044, 2004.
22. Team RDC: R: A Language and Environment for Statistical Computing. Vol. http://www.R-project.org. 2006.
23. Park MY and Hastie T: L1-regularization path algorithms for generalized linear models. J R Stat Soc B 69: 659-677, 2007.
24. Efron B, Hastie T, Johnstone I and Tibshirani R: Least angle regression. Ann Stat 32: 407-499, 2004.
25. Tibshirani R: Regression shrinkage and selection via the lasso. J R Stat Soc B 58: 267-288, 1996.
26. Petricoin EF, Belluco C, Araujo RP and Liotta LA: The blood peptidome: a higher dimension of information content for cancer biomarker discovery. Nat Rev Cancer 6: 961-967, 2006.
27. Diamandis EP: Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 96: 353-356, 2004.
28. Diamandis EP: Serum proteomic profiling by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry for cancer diagnosis: next steps. Cancer Res 66: 5540-5541, 2006.
29. Gao WM, Kuick R, Orchekowski RP, *et al*: Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. BMC Cancer 5: 110, 2005.
30. Maciel CM, Junqueira M, Paschoal ME, Kawamura MT, Duarte RL, Carvalho Mda G and Domont GB: Differential proteomic serum pattern of low molecular weight proteins expressed by adenocarcinoma lung cancer patients. J Exp Ther Oncol 5: 31-38, 2005.
31. Yang SY, Xiao XY, Zhang WG, *et al*: Application of serum SELDI proteomic patterns in diagnosis of lung cancer. BMC Cancer 5: 83, 2005.
32. Yildiz PB, Shyr Y, Rahman JS, *et al*: Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. J Thorac Oncol 2: 893-901, 2007.
33. Cho WC, Yip TT, Yip C, *et al*: Identification of serum amyloid a protein as a potentially useful biomarker to monitor relapse of nasopharyngeal cancer by serum proteomic profiling. Clin Cancer Res 10: 43-52, 2004.
34. Zhukov TA, Johanson RA, Cantor AB, Clark RA and Tockman MS: Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. Lung Cancer 40: 267-279, 2003.
35. Ali IU, Xiao Z, Malone W, *et al*: Plasma proteomic profiling: search for lung cancer diagnostic and early detection markers. Oncol Rep 15: 1367-1372, 2006.
36. Agranoff D, Fernandez-Reyes D, Papadopoulos MC, *et al*: Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. Lancet 368: 1012-1021, 2006.

37. Findeisen P, Zapatka M, Peccerella T, Matzk H, Neumaier M, Schadendorf D and Ugurel S: Serum amyloid A as a prognostic marker in melanoma identified by proteomic profiling. J Clin Oncol 27: 2199-2208, 2009.
38. Herbst RS, Heymach JV and Lippman SM: Lung cancer. N Engl J Med 359: 1367-1380, 2008.
39. Yanagisawa K, Shyr Y, Xu BJ, *et al*: Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 362: 433-439, 2003.
40. Yanagisawa K, Tomida S, Shimada Y, Yatabe Y, Mitsudomi T and Takahashi T: A 25-signal proteomic signature and outcome for patients with resected non-small-cell lung cancer. J Natl Cancer Inst 99: 858-867, 2007.
41. Taguchi F, Solomon B, Gregorc V, *et al*: Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. J Natl Cancer Inst 99: 838-846, 2007.
42. Jacot W, Lhermitte L, Dossat N, *et al*: Serum proteomic profiling of lung cancer in high-risk groups and determination of clinical outcomes. J Thorac Oncol 3: 840-850, 2008.
43. Patz EF Jr, Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, and Herndon J II: Panel of serum biomarkers for the diagnosis of lung cancer. J Clin Oncol 25: 5578-5583, 2007.