Cancer Biomarker Discovery via Targeted Profiling of Multiclass Tumor Tissue-Derived Proteomes

Longhai Zhou · Ming Cai · Xuefeng Bruce Ling · Qiang Wang · Kenneth Lau · Jiagang Jack Zhao · James Schilling · Liangbiao Chen

Abstract

Introduction Tumor-derived proteins and naturally occurring peptides represent a rich source of potential cancer markers for multiclass cancer distinction.

Materials and Methods In this study, proteomes/peptidomes derived from primary colon cancer, kidney cancer, liver cancer, and glioblastoma were analyzed by liquid chromatography coupled with mass spectrometry to identify multiclass cancer discriminative protein and peptide candidates. Spectral counting and peptidomic analyses found two biomarker panels, one with 12 proteins and the other with 53 peptides, both capable of multiclass cancer detection and classification.

Results and Discussion Shed from tumor tissues through apoptosis/necrosis, cell secretion, or tumor-specific degradation of extracellular matrix proteins, these proteins/peptides are likely to enter into circulation and, therefore, have the potential to be configured into practical serological diagnostic and prognostic utilities.

Keywords Cancer classification · Proteomics · Biomarker panel

Longhai Zhou, Ming Cai, and Xuefeng Bruce Ling contributed equally to this study as first authors.

Xuefeng Bruce Ling and Liangbiao Chen contributed equally to this study as senior authors.

L. Zhou · L. Chen (✉)
Key Laboratory of Molecular and Developmental Biology,
Institute of Genetics and Developmental Biology,
Chinese Academy of Sciences,
Beijing 100101, China
e-mail: lbchen@genetics.ac.cn

M. Cai · Q. Wang
Department of Urinary Surgery, Organ Transplantation Center,
Second Affiliated Hospital of Chinese PLA General Hospital,
Beijing 100091, China

X. B. Ling (✉) · K. Lau · J. Schilling
Stanford University,
Stanford, CA 94305, USA
e-mail: xuefeng_ling@yahoo.com

J. J. Zhao
Stemnex Inc.,
Menlo Park, CA 94025, USA

Introduction

Due to their common availability and the potential of immediate translation of found biomarkers into serological clinical utilities, plasma and serum from cancer patients have been extensively analyzed by novel proteomic profiling technology leading to many candidate markers. However, few of these candidate biomarkers have been validated with the desired sensitivity and specificity to allow early cancer detection and prognostic of clinic outcomes [1]. The best single cancer biomarkers may have already been discovered, so future cancer biomarker utilities most likely will be biomarker panels with multiple, less sensitive, and specific biomarkers, in combination with biostatistics modeling, to devise predictive algorithms to achieve required sensitivity and specificity for cancer diagnosis and prognosis [2].

We have followed a cancer tissue-targeted proteomic approach [3], aiming to ultimately discover low, abundant, while cancer-specific proteins in plasma or serum. The rationale is that tumor-derived proteins, secreted by cancer cells or shed from the cancer microenvironment, can

eventually enter the bloodstream and that these proteins' serological abundance could be assessed in combination with a biostatistics model for cancer prediction. We reason that, targeted analysis of these proteins, trapped in the source tumor tissues just prior to their release in circulation, can result in the discovery of even lower abundance, tissue specific, and circulating biomarkers. Therefore, conditioned media derived from primary tumor tissues, expected to enrich with these potential biomarkers, are targeted for proteomic profiling analysis.

In this study, we assayed condition media derived from four tumor types—colon cancer (CC), kidney cancer (KC), liver cancer (LC), and glioblastoma (BC). The mass spectrometric-based proteomic profiling analysis found a biomarker panel of 12 proteins having differential abundance between different tumor types. In addition, we performed comprehensive peptidomic analysis to overlay and compare all the mass spectrometric spectra from various tumor samples for differential tumor-derived marker signals. We have identified a panel of 53 biomarkers, including both tryptic peptides and non-tryptic peptides, capable of discriminating between these cancer types.

Materials and Methods

Samples Tissue specimens were obtained with the approval of the Committee on the Ethical of Research involving Human Subjects from the Affiliated Hospital of Chinese PLA General Hospital. The total 20-case samples contained three colon cancer, six kidney cancer, three liver cancer, four glioblastoma, two ureteral cancer patients, and normal organ samples from one kidney and one liver cancer patients and were histologically confirmed by two independent pathologists. Following surgical resection, tumor tissues were cut into small pieces with sterile scissors and rinsed with PBS several times and placed in 50-ml conical tubes containing defined medium [Dulbecco's modified Eagle medium (DMEM/F12) supplemented with growth factor cocktail, which includes basic FGF 20 ng/ml, EGF 20 ng/ml, insulin 7 μg/ml, and transferin 15 μg/ml, plus penicillin 500 units/ml and streptomycin 500 μg/ml] overnight at 4°C. Following centrifugation for 10 min at 2,000 rpm, the tissue media were desalted through the PD-10 column (GE health care) pre-equilibrated with 0.01% NH$_4$OH, then lyophilized and stored at −80°C.

Preparation of Tumor-Derived Protein Samples The frozen pellets were sonicated and dissolved in 7 M urea and 2 M thiourea and 25 mM ammonium bicarbonate for 2 h. The resulting protein extracts were desalted using Pierce zeba desalt spin columns. Each sample's total protein content was quantified by Pierce BCA protein assay reagent. The desalted samples were diluted with 25 mM ammonium bicarbonate to the same protein concentration 0.5 μg/μl. For reduction, 50 μg protein of each sample was incubated with 10 μl 5 mM DTT at 50°C for 30 min.

LCMS and MSMS analysis For alkylation, iodoacetamide was added to a final concentration of 15 mM. After incubating at room temperature in the dark for 30 min, 1 μg trypsin was added to each sample to digest at 37°C overnight; 1.5 μl 50% TFA in water was added to terminate the reaction. The total volumes of the digests were subsequently dried to ~70 μl in a SpeedVac. Trypsin-digested samples were diluted, 1:10 in 0.1% *v/v* formic acid, and loaded online to an analytical C18 column (75 μm, 12 cm). Peptides from each tumor sample were eluted using a linear gradient of H$_2$O/CH$_3$CN (95:5, 0.1% formic acid buffer A) to H$_2$O/CH$_3$CN (70:30, 0.1% formic acid buffer B) at 300 nl/min over 70 min using a 2D Eskigent nano HPLC, Spark autosampler system. Each tumor sample's full mass spectrometry (MS) scan (from 400 to 1,600 *m/z*) acquired on an LTQ FTMS (Thermo, San Jose, CA, USA) was followed by five MS/MS events using data-dependent acquisition where the first most intense ion from a given MS scan was subjected to CID followed by the second to fifth most intense ions. Protein identification was performed by searching Swiss-Prot protein database using Thermal BioWorks™ software and SEQUEST® algorithm (Thermo, San Jose, CA, USA). Peptide identifications were considered acceptable if they passed the thresholds determined acceptable for human plasma by Qian et al. [4] and passed an additional filter of a PeptideProphet score of at least 0.7 [5]. The PeptideProphet score is representative of the quality of the SEQUEST™ identification and is based on a combination of XCorr, delCn, Sp, and a parameter that measures the probability that the identification occurred by random chance. PeptideProphet scores are normalized to a 0 to 1 scale, with 1 being the highest confidence value.

Spectral Counting Analysis Quantification of proteins in different samples was done by means of spectral counting using Scaffold software (Proteome Software, Portland, OR, USA). From the MS/MS protein identifications, a separate list of proteins was created for each sample, and the lists were then compared to find differential expressed proteins. For any given protein, the relative abundance between samples was estimated by the comparative analysis of the normalized spectrum counts of the identified tryptic peptides.

Peptidomic Data Analysis Our approach, which is commonly referred to as ion mapping [6, 7], first selects biomarker candidate MS peaks on the basis of discriminant

analysis and then targets them for MS/MS sequencing analysis to obtain protein identification. The peak finding and comparative analysis were performed as described in previous work [8, 9].

Results and Discussion

In this study (outlined in Supplementary Fig. 1), we collected a total of 16 primary tumor samples from three colon cancer, six kidney cancer, three liver cancer, four glioblastoma cancer patients, and tumor adjacent tissue counterparts from one kidney and one liver cancer patients. To extract tumor-derived proteins/peptides trapped in the tumor tissue, tissue specimens were rinsed and cut into small pieces and kept at 4°C overnight in defined medium such that tissue-derived and extracellular matrix-derived proteins and peptides can be released. After tryptic digestion, the peptides from the conditioned media were fractionated through C_{18} reverse-phase HPLC and later analyzed by an LTQ FT MS.

A total of 1,807 proteins were identified from control and tumor samples after SEQUEST™ analysis against the Swiss-Prot human database. Protein identification results, e.g., protein hits of different tryptic peptides of the same protein precursor, of all tumor samples were integrated using probabilistic protein identification algorithms implemented in the Scaffold software. Spectrum counts were analyzed from the number of MS/MS spectra identified corresponding to each protein normalized to account for protein length or expected number of tryptic peptides. For any given protein, the relative abundance between samples was estimated by the comparative analysis of the normalized spectrum counts. The box-whisker graphs in Fig. 1 illustrate the spread of the distribution of the spectrum

counts for each identified protein, using a "box" (25~75%) and "whiskers" to break down data by percentile. The results show that 12 proteins, including albumin (ALB), serotransferrin (TF), aplipoprotein A1 (APO A1), Vimentin (VIM), immunoglobulin heavy constant gamma 1(IGHG1), glial fibrillary acidic protein (GFAP), alpha 1 antitrypsin (A1AT), hemoglobin beta (HBB), orosomucoid 1 (AGP1, alias ORM1), pyruvate kinase type M2 (PKM2, alias M2-PK), keratin 8 (KRT8), and keratin 19 (KRT19), appeared to have differential abundance in colon, kidney, liver, and brain tumors. The remaining 1,795 proteins' abundance was largely undifferentiated in the compared tumors. Examination of the 12 differentially expressed proteins found that they can be divided into five groups of expression patterns (Fig. 1). In group1, GFAP was the only protein of higher abundance in brain tumor than that in colon, kidney, and liver tumors. In contrast, the remaining 11 proteins were all of lower abundance in brain than those in the other three tumors. PKM2, KRT8, and KRT19 (group 2) were found to be highly expressed in colon cancer and were largely unexpressed in kidney, liver, and brain tumor types. HBB and APO A1 (group 3) had more abundance in kidney tumor than in colon, liver, and brain tumor types. VIM, AGP1, and A1AT (group 4) were found to be more expressed in colon tumor, then kidney tumor, then liver tumor, and least in brain tumor. TF and ALB (group 5) were more abundant in colon and liver tumor than those in kidney and brain tumors.

Literature review has found that most, if not all, of these 12 proteins can be readily detected in the circulation and have been previously found by others to have diagnostic or prognostic values in various assayed tumors. In renal cell carcinoma, VIM staining has been identified as an independent predictor of poor prognosis, and increased VIM staining correlated with worse survival [10]. ALB, GFAP, and VIM have been discovered by proteomic
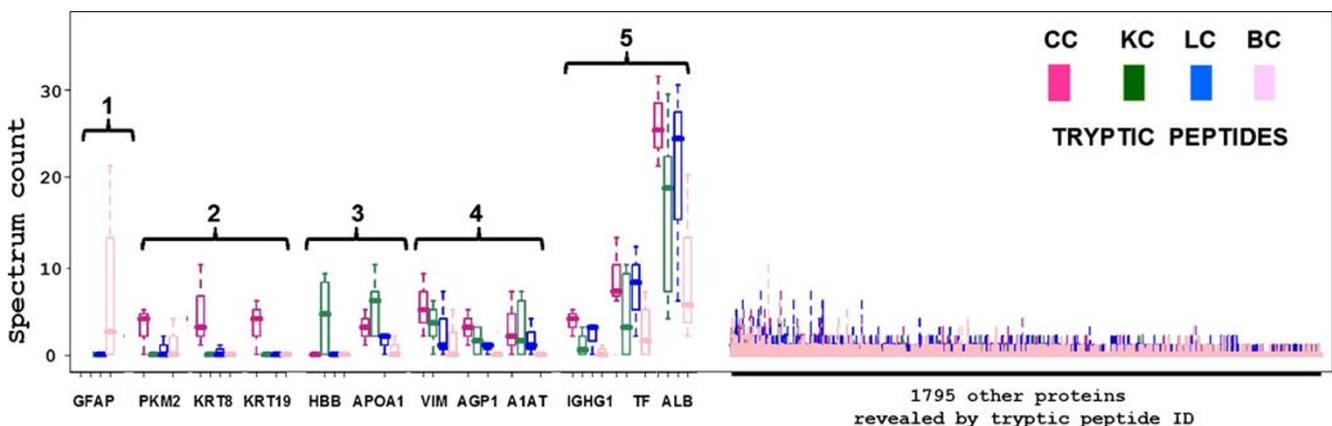
Fig. 1 Mass spectrometric spectral counting analysis qualitatively evaluates the relative abundance of the tumor-derived proteins in colon (*CC*), kidney (*KC*), liver (*LC*), and brain (*BC*) tumors

profiling as molecular indicators of diagnostic or prognostic value for gliomas [11]. A1AT levels have been found to increase significantly in the sera of patients with gastrointestinal cancers, correlating with the stage and severity of the gastric and colorectal cancers [12]. IGHG1 has been shown to be one of the serum "factors" secreted by the epithelial cancers in lung, liver, colon, and breast to protect the neoplastic cells from the lymphocyte reactivity [13]. PKM2 has been shown as a very useful biomarker for early detection of various tumors [14]. Examination of KRT 8, 18, and 19 revealed a consistent pattern of expression with respect to tumor grade. mRNA expression for KRT 8 was significantly higher in node-positive compared with node-negative disease stages [15]. Serum AGP1 profiles can provide prognostic information in patients with glioblastoma multiforme [16]. Very few studies examined HBB gene expression in tumor, and our finding that it is up-regulated in kidney tumor is intriguing. Nevertheless, interrogation of the NCBI GEO database (dataset record GDS505) [17] revealed that HBB, indeed, at the gene expression level, up-regulated in renal clear cell carcinoma (Supplementary Fig. 2). Therefore, our findings of these 12 proteins' differential abundance in various cancers are consistent

with previous analyses that these 12 proteins, to be validated, indeed have differential significance in various tumors and are all of potential serological biomarker candidates. However, it is important to point out that the circulating proteins found by others, albeit sharing the same protein precursors with the ones we discovered in the tumor conditioned media, most likely are protein isoforms and have been post-translationally processed further by clipping/cleavage and modifications such as glycosylation. The post-translational modification or proteolytic enzymes may be themselves biomarkers indicative of cancer pathophysiology [18]. Immunological, enzymatic, or mass spectrometry-based methods are needed to characterize these potentially novel biomarker activities and evaluate their potential clinical utilities in cancer management. In this study, our current focus is the analysis of the relative protein quantifications between cancer types and the potential clinical utilities of the differential abundance for diagnostic and prognostic applications. The characterizations of protein modification activities, suggested by the literature review, will be followed up after validation studies.

In addition to the identity-based spectral counting analysis, we have also performed a comprehensive analysis
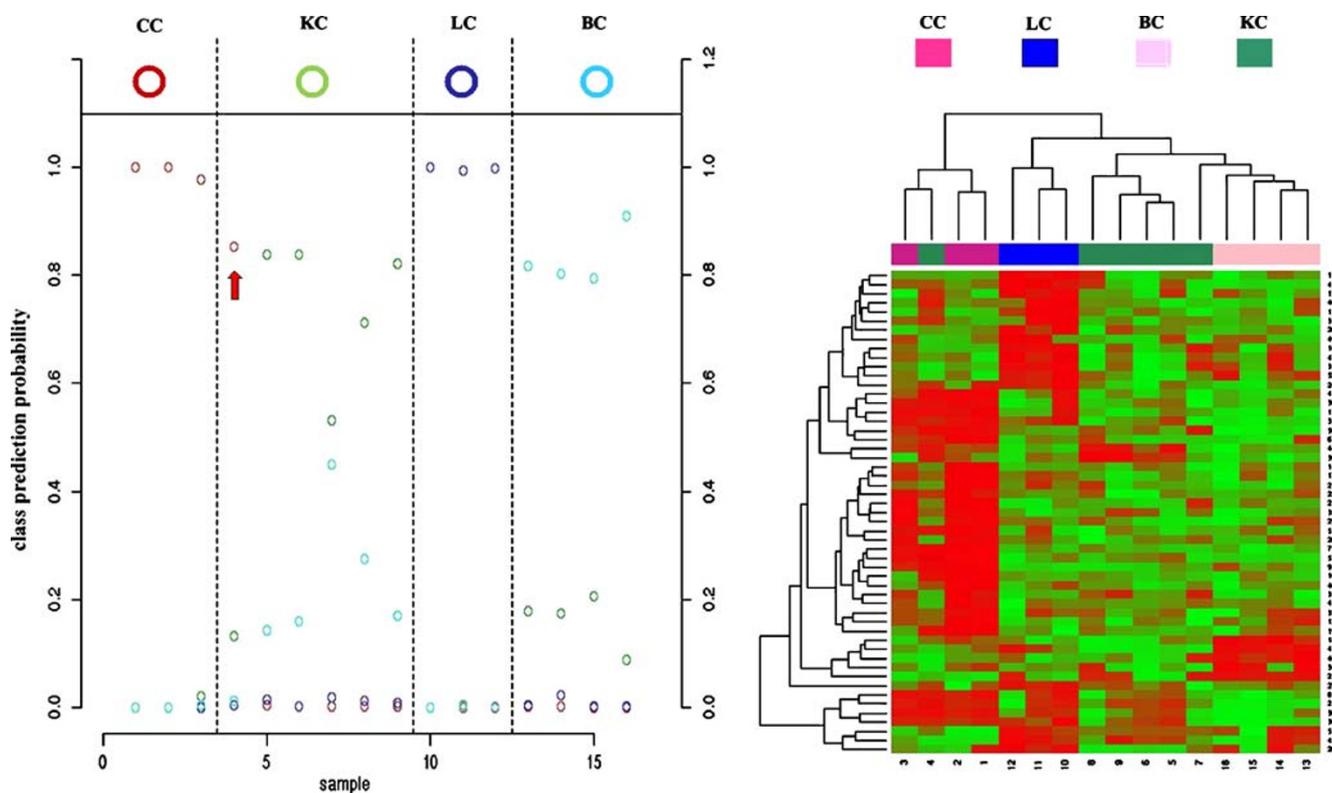
Fig. 2 Analysis of the 53-peptide biomarker panel. *Left panel* With the 53-feature peptide biomarker panel, the predicted discriminant probabilities of the CC, KC, LC, and BC classes for each sample were calculated from the linear discriminant analysis. The maximum estimated probability for each of the wrongly classified samples is marked with a *red arrow*. *Right panel* Unsupervised two dimensional clustering of all CC, KC, LC, and BC samples and the corresponding 53 peptide biomarkers. Heat map reveals relative abundance of these 53 peptide biomarkers in the CC, KC, LC, and BC tumor categories

comparing all MS scans to discover differential tryptic and non-tryptic peptide biomarkers. The non-tryptic peptides are likely to be the result of the tumor-specific degradation of extracellular matrix proteins by proteases and exopeptidase released from cancer cells [18]. A total of 28,000 unique peak features with distinct m/z and HPLC fraction have resolved. The samples were utilized as a training set (CC, $n=3$; KC, $n=6$; LC, $n=3$; BC, $n=4$) for predictor discovery by a nearest shrunken centroid algorithm [19] with all the features in the data set. As shown in Supplementary Fig. 3, internal 4-fold cross-validation and linear discriminant analysis selected a 53-feature panel as the peptide biomarker panel with predictive utility for follow-up analysis.

The peptide panel included both tryptic peptides: a disintegrin and metalloproteinase domain 8 (ADAM8), orosomucoid 2 (AGP2, alias ORM2), immunoglobulin kappa constant (IGKC), MKI67 (FHA domain) interacting nucleolar phosphoprotein (MKI67IP), tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta polypeptide (YWHAZ), and non-tryptic peptides: ankyrin repeat and SOCS box-containing 13 (ASB13), Cyclin-J, glycoprotein Ib (platelet) alpha polypeptide (GP1BA), immunoglobulin superfamily, member 8 (IGSF8), RUN and FYVE domain containing 4 (RUFY4), transient receptor potential cation channel subfamily M member 6 (TRPM6), and zinc finger and SCAN domain containing 4 (ZSCAN4), capable of discriminating between the colon, kidney, liver, and brain tumors. Figure 2 (left panel) displays the linear discriminative analysis predictive probabilities where tumor samples have robust separation between the highest and next highest probability, demonstrating that most of the samples were correctly classified into CC, KC, LC, and BC tumor categories using the 53-peptide panel. With the maximum estimated probability marked with a red arrow, only one of the KC samples was wrongly classified. Consistent with these findings, unsupervised clustering (Fig. 2, right panel) based upon the 53-peptide biomarker panel was able to largely cluster, according to their diagnosis, and only one of the brain tumor was clustered within the colon cancer samples.

Significant degree of shared gene expression exists between tumors and their normal tissue counterparts. Expression analysis comparing different tumors may risk uncovering tissue differentiation markers indicative of lineage differentiation, which is irrelevant to oncogenic process. This prompted us to investigate whether our 53-peptide biomarker panel indeed captured the "molecular portraits" of the assayed cancers rather than those of their paired normal tissues. We have performed two comparisons of either the kidney or liver tumor tissues and the adjacent normal tissue counterparts isolated from the same patient. In both kidney and liver cases, samples of the tumor or adjacent normal tissue counterpart from the same patient (Fig. 3) cluster together. However, in both cases of the kidney and liver, expression profiles between the tumor and adjacent normal tissue counterpart were clearly different. Therefore, our 53-peptide biomarker panel can discriminate and cluster different tumor types in both supervised and unsupervised clustering analyses and can delineate tumor from the adjacent normal tissue. However, it is important to interpret data with caution due to the small sample size of various assayed tumor tissues. Prospective validation studies with sufficient sample size of enough analytic power are essential to substantiate these peptides' role in cancer diagnosis.

Peptides within the 53-peptide biomarker panel have been subjected to extensive protein identification efforts via LTQ FT MS/MS and database searches upon both the tryptic and non-tryptic peptide fingerprinting analyses. Of the 53 peptide features (Supplementary Table 1), 18 were positively identified where five peptides are non-tryptic and 13 are tryptic. Tryptic peptides of ALB, APOA1, and KRT8 were found in the 53-peptide biomarker panel, where the quantification analysis results of these peptides are in line with those obtained from previous spectral counting
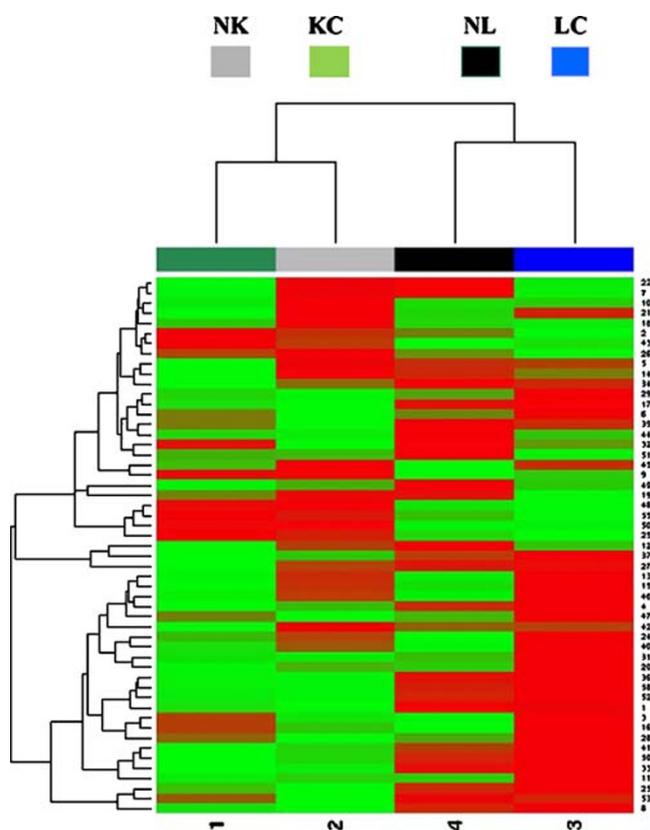
Fig. 3 Comparative analysis and heat map plot of the relative abundance of the 53 peptide biomarkers in either the kidney or liver tumor tissue and their corresponding adjacent normal tissue counterpart isolated from the same patient

analysis: Three different tryptic ALB peptides were identified to have higher abundance in colon and liver cancer categories; one tryptic KRT8 peptide was identified to have higher abundance in colon cancer samples; one tryptic APO A1 was identified to have higher abundance in kidney cancer. Tryptic peptides from ADAM8, AGP2, IGKC, KRT18, MKI67IP, and YWHAZ were also found. However, their parent proteins were shown to be undifferentiated by spectral counting analysis. Non-tryptic peptides from ASB13, Cyclin-J, GP1BA, IGSF8, RUFY4, TRPM6, and ZSCAN4 were found to be part of the 53-peptide biomarker panel. Based on our current experience there are three reasons for failure to obtain the peptide sequences of the remaining peptide biomarkers: (1) Peptides are too low in abundance in the original samples for successful MS/MS. (2) Peptides appear to have adequate signals in MS mode but do not produce a sufficient number of product ions in MS/MS to allow definitive protein identification. (3) Peptides produce MS/MS spectra that appear to be adequate but cannot be interpreted by the currently available software. In some instances we are able to solve these by manual interpretation. We have found a significant number of post-translational modifications of tumor-derived peptides that complicate automated spectral interpretation. Future efforts will utilize a two- or three-dimensional HPLC purification prior to MS/MS analysis to increase the sample load and/or the purity of the peptide since peptide ionization efficiency can be related to the purity of the sample.

As shown in Supplementary Fig. 4, we hypothesize that cancer microenvironment, in the similar fashion as that seen in serum [18], can generate and shed naturally occurring but tumor-specific peptides. Comprehensive peptidomic analysis identified a panel of 53 peptide biomarkers, including both tryptic and non-tryptic peptides, capable of discriminating between these tumors. Concerns have been raised regarding the serum naturally occurring peptide biomarker discovery efforts: One major consideration of the serum peptidome discovery [18, 20, 21] is that most, if not all, of the peptide biomarkers are derived from a low number of plasma high-abundance proteins; due to the substantial endogenous endoproteolytic and exoproteolytic enzymatic activity, serum peptidome content can be influenced by sample collection and, therefore, could give rise to artifacts [22]. However, our tumor-derived peptidomic analysis did not suffer from those above issues concerning serum peptidomic analysis. Therefore, we believe that the tumor-derived peptidomic patterns should represent genuine difference between various tumors and their normal tissue counterparts. Future prospective studies of these tumor-derived biomarkers, either by antibody-based or quantitative mass spectrometry-based approach, can optimize them into practical clinical utilities for serological diagnosis and prognosis.

References

1. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst. 2004;96:353–6.
2. Diamandis EP, van der Merwe DE. Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. Clin Cancer Res. 2005;11:963–5.
3. Zhang H, Chan DW. Cancer biomarker discovery in plasma using a tissue-targeted proteomic approach. Cancer Epidemiol Biomarkers Prev. 2007;16:1915–7.
4. Qian WJ, Jacobs JM, Liu T, Camp DG 2nd, Smith RD. Advances and challenges in liquid chromatography–mass spectrometry-based proteomics profiling for clinical applications. Mol Cell Proteomics. 2006;5:1727–44.
5. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002;74:5383–92.
6. Fach EM, Garulacan LA, Gao J, Xiao Q, Storm SM, Dubaquie YP, et al. In vitro biomarker discovery for atherosclerosis by proteomics. Mol Cell Proteomics. 2004;3:1200–10.
7. Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA. Changes in the protein expression of yeast as a function of carbon source. J Proteome Res. 2003;2:643–9.
8. Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics. 2004;20:3034–44.
9. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL Jr, Qu Y, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics. 2003;4:449–63.
10. Lam JS, Leppert JT, Figlin RA, Belldegrun AS. Role of molecular markers in the diagnosis and therapy of renal cell carcinoma. Urology. 2005;66:1–9.
11. Khalil AA. Biomarker discovery: a proteomic approach for brain cancer profiling. Cancer Sci. 2007;98:201–13.
12. Bernacka K, Kuryliszyn-Moskal A, Sierakowski S. The levels of alpha 1-antitrypsin and alpha 1-antichymotrypsin in the sera of patients with gastrointestinal cancers during diagnosis. Cancer. 1988;62:1188–93.
13. Chen Z, Gu J. Immunoglobulin G expression in carcinomas and cancer cell lines. FASEB J. 2007;21:2931–8.
14. Mazurek S, Boschek CB, Hugo F, Eigenbrodt E. Pyruvate kinase type M2 and its role in tumor growth and spreading. Semin Cancer Biol. 2005;15:300–8.
15. Brotherick I, Robson CN, Browell DA, Shenfine J, White MD, Cunliffe WJ, et al. Cytokeratin expression in breast cancer: phenotypic changes associated with disease progression. Cytometry. 1998;32:301–8.
16. Matsuura H, Nakazawa S. Prognostic significance of serum alpha 1-acid glycoprotein in patients with glioblastoma multiforme: a preliminary communication. J Neurol Neurosurg Psychiatry. 1985;48:835–7.
17. Lenburg ME, Liou LS, Gerry NP, Frampton GM, Cohen HT, Christman MF. Previously unidentified changes in renal cell

carcinoma gene expression identified by parametric analysis of microarray data. BMC Cancer. 2003;3:31.

18. Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, et al. Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest. 2006;116:271–84.

19. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA. 2002;99:6567–72.

20. Villanueva J, Martorella AJ, Lawlor K, Philip J, Fleisher M, Robbins RJ, et al. Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. Mol Cell Proteomics. 2006;5: 1840–52.

21. Villanueva J, Nazarian A, Lawlor K, Yi SS, Robbins RJ, Tempst P. A sequence-specific exopeptidase activity test (SSEAT) for "functional" biomarker discovery. Mol Cell Proteomics. 2008;7: 509–18.

22. Koomen JM, Li D, Xiao LC, Liu TC, Coombes KR, Abbruzzese J, et al. Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. J Proteome Res. 2005;4:972–81.