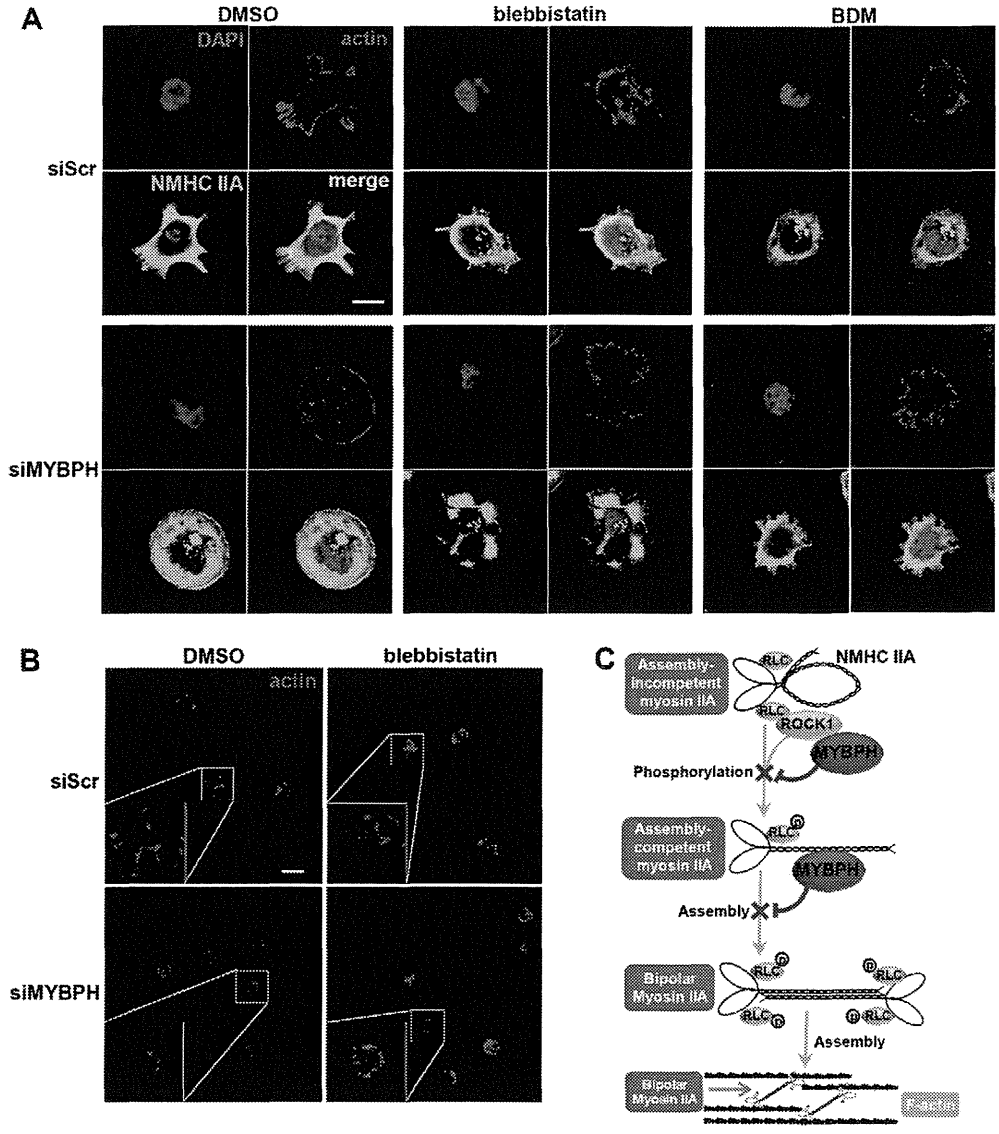

Fig. 4. siMYBPH-induced peripheral actomyosin bundle formation is counteracted by simultaneous treatment with non-muscle myosin inhibitors. (A) Immunofluorescence staining for actin (red) and NMHC IIA (green) in blebbistatin- or BDM-treated NCI-H441 cells. Immunofluorescence staining was performed as previously described [26]. Bar indicates 10 μm. (B) Three-dimensional Matrigel invasion assay in NCI-H441 cell treated with siMYBPH and/or blebbistatin. Three-dimensional Matrigel invasion assays were performed as previously described [26]. White bar indicates 50 μm. (C) Schematic diagram of multifaceted inhibitory roles of MYBPH in actomyosin organization at 2 distinct steps.

to search for MYBPH alterations in cases with similar disease phenotypes without NMHC IIA mutations.

In summary, our results demonstrate that MYBPH inhibits the assembly of NMHC IIA through direct binding to assembly-competent NMHC IIA, suggesting that this activity may in turn contribute to suppression of cancer invasion and metastasis together with its ROCK1 inhibitory function [26]. The dual roles of MYBPH in NM IIA inhibition comprise an intriguing mechanism to impose firm NM IIA inhibition. The present findings also provide clues for better understanding of the molecular mechanisms involved in inhibition of cancer invasion and metastasis by TTF-1 through transcriptional activation of MYBPH, as well as for better prognosis for TTF-1-positive lung adenocarcinoma patients.

Acknowledgments

This work was supported in part by Grants-in-Aid for Scientific Research on Innovative Areas from The Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, as well as a Grant-in-Aid for Young Scientists (B) from the Japan Society for the Promotion of Science (JSPS). Y.H. was supported by a research fellowship of the Foundation for Promotion of Cancer Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2012.10.036.

References

[1] D.A. Lauffenburger, A.F. Horwitz, Cell migration: a physically integrated molecular process, Cell 84 (1996) 359–369.

[2] P. Friedl, K. Wolf, Tumour-cell invasion and migration: diversity and escape mechanisms, Nat. Rev. Cancer 3 (2003) 362–374.

[3] A. Hall, The cytoskeleton and cancer, Cancer Metastasis Rev. 28 (2009) 5–14.

[4] M.A. Conti, R.S. Adelstein, Nonmuscle myosin II moves in new directions, J. Cell Sci. 121 (2008) 11–18.

[5] M. Vicente-Manzanares, X. Ma, R.S. Adelstein, A.R. Horwitz, Non-muscle myosin II takes centre stage in cell adhesion and migration, Nat. Rev. Mol. Cell Biol. 10 (2009) 778–790.
[6] V. Betapudi, L.S. Licate, T.T. Egelhoff, Distinct roles of nonmuscle myosin II isoforms in the regulation of MDA-MB-231 breast cancer cell spreading and migration, Cancer Res. 66 (2006) 4725–4733.
[7] Y. Huang, P. Arora, C.A. McCulloch, W.F. Vogel, The collagen receptor DDR1 regulates cell spreading and motility by associating with myosin IIA, J. Cell Sci. 122 (2009) 1637–1646.
[8] S. Medjkane, C. Perez-Sanchez, C. Gaggioli, E. Sahai, R. Treisman, Myocardin-related transcription factors and SRF are required for cytoskeletal dynamics and experimental metastasis, Nat. Cell. Biol. 11 (2009) 257–268.
[9] S. Etienne-Manneville, A. Hall, Rho GTPases in cell biology, Nature 420 (2002) 629–635.
[10] K. Riento, A.J. Ridley, Rocks: multifunctional kinases in cell behaviour, Nat. Rev. Mol. Cell Biol. 4 (2003) 446–456.
[11] K. Itoh, K. Yoshioka, H. Akedo, M. Uehata, T. Ishizaki, S. Narumiya, An essential part for Rho-associated kinase in the transcellular invasion of tumor cells, Nat. Med. 5 (1999) 221–225.
[12] E. Sahai, C.J. Marshall, ROCK and Dia have opposing effects on adherens junctions downstream of Rho, Nat. Cell Biol. 4 (2002) 408–415.
[13] S. Wilkinson, H.F. Paterson, C.J. Marshall, Cdc42-MRCK and Rho-ROCK signalling cooperate in myosin phosphorylation and cell invasion, Nat. Cell Biol. 7 (2005) 255–261.
[14] C.C. Wong, C.M. Wong, F.C. Ko, L.K. Chan, Y.P. Ching, J.W. Yam, I.O. Ng, Deleted in liver cancer 1 (DLC1) negatively regulates Rho/ROCK/MLC pathway in hepatocellular carcinoma, PLoS One 3 (2008) e2779.
[15] N.G. Dulyaninova, V.N. Malashkevich, S.C. Almo, A.R. Bresnick, Regulation of myosin-IIA assembly and Mts1 binding by heavy chain phosphorylation, Biochemistry 44 (2005) 6867–6876.
[16] Z.H. Li, A.R. Bresnick, The S100A4 metastasis factor regulates cellular motility via a direct interaction with myosin-IIA, Cancer Res. 66 (2006) 5173–5180.
[17] S. Kimura, Y. Hara, T. Pineau, P. Fernandez-Salguero, C.H. Fox, J.M. Ward, F.J. Gonzalez, The T/ebp null mouse: thyroid-specific enhancer-binding protein is essential for the organogenesis of the thyroid, lung, ventral forebrain, and pituitary, Genes Dev. 10 (1996) 60–69.
[18] Y. Yatabe, T. Mitsudomi, T. Takahashi, TTF-1 expression in pulmonary adenocarcinomas, Am. J. Surg. Pathol. 26 (2002) 767–773.
[19] T. Takeuchi, S. Tomida, Y. Yatabe, T. Kosaka, H. Osada, K. Yanagisawa, T. Mitsudomi, T. Takahashi, Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors, J. Clin. Oncol. 24 (2006) 1679–1688.
[20] H. Tanaka, K. Yanagisawa, K. Shinjo, A. Taguchi, K. Maeno, S. Tomida, Y. Shimada, H. Osada, T. Kosaka, H. Matsubara, T. Mitsudomi, Y. Sekido, M. Tanimoto, Y. Yatabe, T. Takahashi, Lineage-specific dependency of lung adenocarcinomas on the lung development regulator TTF-1, Cancer Res. 67 (2007) 6007–6011.
[21] J. Kendall, Q. Liu, A. Bakleh, A. Krasnitz, K.C. Nguyen, B. Lakshmi, W.L. Gerald, S. Powers, D. Mu, Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer, Proc. Natl. Acad. Sci. USA 104 (2007) 16663–16668.
[22] B.A. Weir, M.S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W.M. Lin, M.A. Province, A. Kraja, L.A. Johnson, K. Shah, M. Sato, R.K. Thomas, J.A. Barletta, I.B. Borecki, S. Broderick, A.C. Chang, D.Y. Chiang, L.R. Chirieac, J. Cho, Y. Fujii, A.F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B.E. Johnson, M.G. Kris, A. Lash, L. Lin, N. Lindeman, E.R. Mardis, J.D. McPherson, J.D. Minna, M.B. Morgan, M. Nadel, M.B. Orringer, J.R. Osborne, B. Ozenberger, A.H. Ramos, J. Robinson, J.A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M.R. Spitz, M.S. Tsao, D. Twomey, R.G. Verhaak, G.M. Weinstock, D.A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M.F. Zakowski, Q. Zhang, D.G. Beer, Wistuba II, M.A. Watson, L.A. Garraway, M. Ladanyi, W.D. Travis, W. Pao, M.A. Rubin, S.B. Gabriel, R.A. Gibbs, H.E. Varmus, R.K. Wilson, E.S. Lander, M. Meyerson, Characterizing the cancer genome in lung adenocarcinoma, Nature 450 (2007) 893–898.
[23] K.A. Kwei, Y.H. Kim, L. Girard, J. Kao, M. Pacyna-Gengelbach, K. Salari, J. Lee, Y.L. Choi, M. Sato, P. Wang, T. Hernandez-Boussard, A.F. Gazdar, I. Petersen, J.D. Minna, J.R. Pollack, Genomic profiling identifies TITF1 as a lineage-specific oncogene amplified in lung cancer, Oncogene 27 (2008) 3635–3640.
[24] T. Yamaguchi, K. Yanagisawa, R. Sugiyama, Y. Hosono, Y. Shimada, C. Arima, S. Kato, S. Tomida, M. Suzuki, H. Osada, T. Takahashi, NKX2-1/TITF1/TTF-1-Induced ROR1 is required to sustain EGFR survival signaling in lung adenocarcinoma, Cancer Cell 21 (2012) 348–361.
[25] V.K. Anagnostou, K.N. Syrigos, G. Bepler, R.J. Homer, D.L. Rimm, Thyroid transcription factor 1 is an independent prognostic factor for patients with stage I lung adenocarcinoma, J. Clin. Oncol. 27 (2009) 271–278.
[26] Y. Hosono, T. Yamaguchi, E. Mizutani, K. Yanagisawa, C. Arima, S. Tomida, Y. Shimada, M. Hiraoka, S. Kato, K. Yokoi, M. Suzuki, T. Takahashi, MYBPH, a transcriptional target of TTF-1, inhibits ROCK1, and reduces cell motility and metastasis, EMBO J. 31 (2012) 481–493.
[27] A.F. Straight, A. Cheung, J. Limouze, I. Chen, N.J. Westwood, J.R. Sellers, T.J. Mitchison, Dissecting temporal and spatial control of cytokinesis with a myosin II inhibitor, Science 299 (2003) 1743–1747.
[28] L.P. Cramer, T.J. Mitchison, Myosin is involved in postmitotic cell spreading, J. Cell Biol. 131 (1995) 179–189.
[29] E. Flashman, C. Redwood, J. Moolman-Smook, H. Watkins, Cardiac myosin binding protein C: its role in physiology and disease, Circ. Res. 94 (2004) 1279–1289.
[30] C. Gaggioli, S. Hooper, C. Hidalgo-Carcedo, R. Grosse, J.F. Marshall, K. Harrington, E. Sahai, Fibroblast-led collective invasion of carcinoma cells with differing roles for RhoGTPases in leading and following cells, Nat. Cell Biol. 9 (2007) 1392–1400.
[31] C. Hidalgo-Carcedo, S. Hooper, S.I. Chaudhry, P. Williamson, K. Harrington, B. Leitinger, E. Sahai, Collective cell migration requires suppression of actomyosin at cell–cell contacts mediated by DDR1 and the cell polarity regulators Par3 and Par6, Nat. Cell Biol. 13 (2011) 49–58.
[32] K.M. Trybus, S. Lowey, Conformational states of smooth muscle myosin. Effects of light chain phosphorylation and ionic strength, J. Biol. Chem. 259 (1984) 8564–8571.
[33] E.J. Kim, D.M. Helfman, Characterization of the metastasis-associated protein, S100A4. Roles of calcium binding and dimerization in cellular localization and interaction with myosin, J. Biol. Chem. 278 (2003) 30063–30073.
[34] J.D. Franke, F. Dong, W.L. Rickoll, M.J. Kelley, D.P. Kiehart, Rod mutations associated with MYH9-related disorders disrupt nonmuscle myosin-IIA assembly, Blood 105 (2005) 161–169.
[35] Y. Zhang, M.A. Conti, D. Malide, F. Dong, A. Wang, Y.A. Shmist, C. Liu, P. Zerfas, M.P. Daniels, C.C. Chan, E. Kozin, B. Kachar, M.J. Kelley, J.B. Kopp, R.S. Adelstein, Mouse models of MYH9-related disease: mutations in nonmuscle myosin II-A, Blood 119 (2012) 238–250.

Hindawi Publishing Corporation
International Journal of Proteomics
Volume 2012, Article ID 510397, 11 pages
doi:10.1155/2012/510397

Research Article

Seven-Signal Proteomic Signature for Detection of Operable Pancreatic Ductal Adenocarcinoma and Their Discrimination from Autoimmune Pancreatitis

Kiyoshi Yanagisawa,[1,2] Shuta Tomida,[2] Keitaro Matsuo,[3] Chinatsu Arima,[2] Miyoko Kusumegi,[4] Yukihiro Yokoyama,[5] Shigeru B. H. Ko,[6] Nobumasa Mizuno,[7] Takeo Kawahara,[5] Yoko Kuroyanagi,[8] Toshiyuki Takeuchi,[8] Hidemi Goto,[6] Kenji Yamao,[7] Masato Nagino,[5] Kazuo Tajima,[3] and Takashi Takahashi[2]

[1] *Institute for Advanced Research, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan*
[2] *Division of Molecular Carcinogenesis, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, Nagoya 466-8550, Japan*
[3] *Division of Epidemiology and Prevention, Aichi Cancer Center, Nagoya 464-8681, Japan*
[4] *Division of Research and Development, Oncomics Co., Ltd., Nagoya 464-0858, Japan*
[5] *Division of Surgical Oncology, Department of Surgery, Nagoya University Hospital, Nagoya 466-8550, Japan*
[6] *Department of Gastroenterology, Nagoya University Hospital, Nagoya 466-8550, Japan*
[7] *Department of Gastroenterology, Aichi Cancer Center, Nagoya 464-8681, Japan*
[8] *Division of Research and Development, Oncomics Co., Ltd, Nagoya 464-0858, Japan*

Correspondence should be addressed to Kiyoshi Yanagisawa, kyana@med.nagoya-u.ac.jp

Received 27 January 2012; Accepted 9 March 2012

Academic Editor: Visith Thongboonkerd

There is urgent need for biomarkers that provide early detection of pancreatic ductal adenocarcinoma (PDAC) as well as discrimination of autoimmune pancreatitis, as current clinical approaches are not suitably accurate for precise diagnosis. We used mass spectrometry to analyze protein profiles of more than 300 plasma specimens obtained from PDAC, noncancerous pancreatic diseases including autoimmune pancreatitis patients and healthy subjects. We obtained 1063 proteomic signals from 160 plasma samples in the training cohort. A proteomic signature consisting of 7 mass spectrometry signals was used for construction of a proteomic model for detection of PDAC patients. Using the test cohort, we confirmed that this proteomic model had discrimination power equal to that observed with the training cohort. The overall sensitivity and specificity for detection of cancer patients were 82.6% and 90.9%, respectively. Notably, 62.5% of the stage I and II cases were detected by our proteomic model. We also found that 100% of autoimmune pancreatitis patients were correctly assigned as noncancerous individuals. In the present paper, we developed a proteomic model that was shown able to detect early-stage PDAC patients. In addition, our model appeared capable of discriminating patients with autoimmune pancreatitis from those with PDAC.

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the fifth leading cause of cancer death in Japan with more than 24,000 deaths annually [1], while 35,000 deaths each year in the United States are caused by the disease [2]. Long-term survival for PDAC patients remains unsatisfactory, with only 3–5% surviving for more than 5 years after surgical resection, with the remainder succumbing to widespread metastasis or massive local recurrence. Since surgical resection is the only reliable curative treatment, early detection is essential to improve the outcomes of affected individuals. However, the clinical symptoms of PDAC are often unremarkable until advanced stages of the disease, and the anatomic

location of the pancreas deep in the abdomen makes physical detection and imaging approaches difficult. Thus, less than 10% of patients diagnosed with PDAC are eligible for surgical resection [3]. Although serum markers for PDAC including carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9) play important roles in current clinical practice for monitoring progression and treatment response, as well as surveillance for recurrence, these markers are not ideal for cancer screening due to their low specificity and/or sensitivity in early stages of the disease [4–6].

The concept of autoimmune pancreatitis (AIP) is supported by recent advances in elucidating its pathogenesis as a unique systemic disease. AIP has several characteristic features, such as infiltration of CD4-positive T cells and IgG4-positive plasmacytes, irregular narrowing of the pancreatic duct, and diffuse enlargement of the pancreas [7–9]. Although intensive investigations into the pathogenesis of AIP have been conducted, its underlying molecular mechanism remains unclear. The most important and difficult step in diagnosing AIP is to distinguish it from PDAC. Clinical symptoms such as obstructive jaundice are not helpful for discrimination, while IgG4, the most accurate serum marker for AIP, is not adequately specific to exclude the existence of cancer. Furthermore, AIP is sometimes accompanied by PDAC; thus percutaneous or endoscopic biopsy findings are often needed for final diagnosis. Unfortunately, those examinations are invasive for the patient and may fail to detect small regions of cancer cells. As a result, unnecessary surgery because of misdiagnosis performed for AIP patients without cancer or those undergoing treatment for existing cancer is a critical issue in clinical practice. Accordingly, there is urgent need for elucidation of novel biomarker(s) and noninvasive diagnostic strategies useful for early detection of PDAC, as well as discrimination of patients with AIP to improve clinical management and prognosis.

Comprehensive analysis of protein expression patterns in biological materials might improve understanding of the molecular complexities of human diseases [10] and could be useful to detect diagnostic or predictive protein expression patterns that reflect clinical features. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI MS) can profile proteins up to 50 kDa in size in serum, tissues, and other various clinical specimens. Protein profiles obtained may contain thousands of data points and provide proteomic signatures that allow detection of patients with various diseases [11, 12]. We previously employed MALDI MS for expression profiling of proteins in human lung cancer specimens and found that the resultant proteomic patterns could predict various clinical features, as well as the potential of recurrence in stage I lung cancer patients [13, 14].

In the present study, protein expression profiling with MALDI MS was conducted to identify proteomic patterns in plasma samples for discrimination of PDAC from AIP as well as chronic pancreatitis (CP) using 3 independent datasets. We found that a proteomic model consisting of 7 mass spectrometry signals constructed by use of the training cohort could detect 82.6% (38 of 46, 95% CI 68.6–92.2) of known PDAC cases, including 62.5% (5 of 8, 95% CI 24.5–91.5) of the stage I and II cases in the independent test cohort, which successfully confirmed its discrimination power. We further applied our model for discrimination of AIP as well as CP from PDAC and found that it correctly assigned 100% of the AIP and CP patients (19 of 19, 95% CI 82.4–100 and 11 of 11, 95% CI 71.5–100, resp.) as noncancerous. These results indicate that our 7-signal proteomic model may contribute to accurate decisions regarding the therapeutic plan for patients with chronic pancreatic diseases, especially PDAC and AIP.

2. Methods

2.1. Patients and Specimens. Plasma specimens from 96 PDAC patients were obtained from the Department of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan, collected from January 2001 and November 2005. Of those, 80 were randomly assigned to the training set and 16 to the test set. An additional 30 plasma specimens from PDAC patients were obtained from the Department of Surgery, Nagoya University Hospital, Nagoya, Japan, collected from May 2004 to July 2006, and assigned to the test set. Plasma specimens from 147 healthy control subjects were also obtained from the Department of Epidemiology and Prevention, Aichi Cancer Center Research Institute, and used. Of those, 80 were randomly assigned to the training set and 67 to the test set. Plasma specimens from 2 acute pancreatitis, 11 chronic pancreatitis, and 3 autoimmune pancreatitis patients were obtained from the Department of Gastroenterology, Nagoya University Hospital, collected from April 2005 and November 2007, and assigned to the test set. In addition, 16 plasma specimens from autoimmune pancreatitis were obtained from the Department of Gastroenterology, Nagoya University Hospital, collected from September 2003 and August 2009, and assigned to the confirmation set. More detailed information is available in Supplementary Material available on line at doi: 10.1155/2012/510397. The characteristics of the patients and healthy subjects in the training, test, and confirmation cohorts are summarized in Supplementary Table S1, which shows that there were no statistically significant differences in regard to clinicopathologic features among the cohorts. All specimens were processed in the same manner and stored at −80°C within 180 minutes after being collected from the patients and healthy subjects, and not thawed until analysis. Requisite approval from our institutional review boards and written informed consent from all subjects were obtained. One plasma specimen per patient or healthy subject was analyzed, and the training, test, and confirmation datasets were independently analyzed as different batches. Further details are available in supplementary Material.

2.2. Proteomic Analysis. Five microliters of nonpre-treated plasma was mixed with 5 nL drops of an energy absorbing matrix solution (saturated Sinapinic acid in water/acetonitrile/trifluoroacetic acid (500 : 500 : 1, by volume), which allows molecules to be protonated and desorbed from tissue surfaces). Then, 1 μL mixtures were deposited into individual wells of MALDI MS sample plates (PE Biosystems, Foster

City, CA) and dried at room temperature for 5 minutes. Six spots were generated for each plasma-matrix mixture sample and spectra were acquired from all 6 using a 4800 Instrument (Applied Biosystems, Foster City, CA), essentially as described previously [13, 14]. Further details are available in Supplementary Material.

2.3. Statistical Methods. Protein profiles obtained by MALDI MS were analyzed using 3 distinct statistical methods, Fisher's exact test, the Kruskal-Wallis test, and a significance analysis of microarray (SAM) test [15], to investigate MS signals that appeared to differentiate PDAC patients from healthy individuals in the training set. MS signals that met at least 1 of the 3 selection criteria were further analyzed.

To construct a generally applicable proteomic classifier without specifically overfitting it to the training cohort, we used a weighted voting algorithm, a well-established technique for supervised classification, in which each weight value was calculated as the signal-to-noise ratio and a leave-one-out cross-validation strategy was utilized [16].

It is possible that unintended biased resubstitution or partial cross-validation can result in underestimation of the error rate after cross-validation; thus the performance of any class prediction rule is best assessed by applying the rule created by use of 1 dataset (the training set) to an independent dataset (the validation or test set) [17]. In the present study, the proteomic classifier constructed with the training dataset of 160 individuals was validated using a completely independent validation set composed of 145 individuals.

An agglomerative hierarchical clustering algorithm was applied to investigate the pattern among the statistically significant discriminator proteins as well as the biological status with Eisen's software [18].

2.4. Identification of Individual Proteins in the Proteomic Signature. 40 μ of serum samples was pretreated with high abundant protein depletion column (Agilent, Palo Alto, CA) according to manufacturer's instruction. The pretreated serum samples were separated over a polymeric column (Toso, Tokyo, Japan) with a high-performance liquid chromatography (HPLC) pump (Shimadzu, Osaka, Japan) and HPLC fractions were collected every minute for 80 minutes. Each fraction was lyophilized, reconstituted with a 50% acetonitrile in water containing 0.1% trifluoroacetic acid, and analyzed by MALDI mass spectrometry to identify the HPLC fractions that contained proteins corresponding to the peaks in the signature with molecular weights selected by bioinfomatic analysis as candidate molecular markers for the PDAC. The selected fractions were lyophilized and reconstituted with a mixture of 10 μL of 0.4 M ammonium hydrogen carbonate and 5 μL of 45 mM dithiothreitol, and then 10 μL of 100 mM iodoacetamide was added. This mixture was incubated for 4 hours at 37°C with 5 μL of 200 nM mass-grade trypsin (Promega, Madison, WI) to obtain peptides. The peptides were separated and sequenced by a microcapillary reverse-phase column (KYA technologies, Tokyo, Japan) with an HPLC pump (KYA) and MALDI mass spectrometer (Applied Biosystems). These spectra were compared with those in the human databases of the National Center for Biotechnology Information (nonredundant) by use of Mascot version 2.1.0 (Matrix Science Inc., Boston, MA). A minimum of two peptide matches and a positive association between the m/z values detected with MALDI mass spectrometry and the molecular weight of the intact protein (including posttranslational modifications) were required for protein identification.

3. Results

3.1. Protein Expression Profiling in the Training Cohort. We obtained protein expression profiles for the 160 human plasma specimens obtained from 80 PDAC patients and 80 healthy subjects at Aichi Cancer Center (Figure 1(a)) and Supplementary Table S1) using MALDI MS. Spectra were obtained from 6 replicates of single plasma specimens. MarkerView (Applied Biosystems) and custom software were used to bin the peaks across the spectra obtained from 960 samples, and then we calculate the average intensity of each signal individually among the 160 cases. As a result, we obtained expression profiles containing 1063 distinct proteomic signals. To extract a proteomic signature able to discriminate PDAC patients from healthy individuals, we compared MS signals from the 80 healthy subjects and 80 PDAC patients using our statistical selection criteria (signals met at least 2 of the following criteria: P value corrected with Bonferroni was less than 0.05 in Fisher's exact test and Kruskal-Wallis test, and FDR < 0.1% for SAM). As a result, 134 MS signals were found to be differentially expressed. Agglomerative hierarchical clustering analysis using the identified proteomic signature showed a clear separation of plasma specimens from PDAC patients as compared to those from healthy individuals (Figure 1(b)), which confirmed that the selected MS signals were informative for discrimination of PDAC cases from healthy individuals. The left branch mostly consisted of PDAC cases (81.3%, 65 of 80 cases, 95% CI 71.0–89.1), whereas the right branch consisted of healthy subjects (78.8%, 63 of 80 cases, 95% CI 68.2–87.1). Next, we investigated whether our proteomic prediction model could best distinguish noncancerous individuals from cancer patients. For this purpose, the 134 selected MS signals, which were informative for discrimination, were further ranked according to the SAM and weighted-voting proteomic discriminatory models were constructed using increasing numbers of the differentially expressed proteomic signals (up to 134), for which learning errors were calculated by leave-one-out cross-validation (Figure 2(a)). This cross-validation analysis showed that the use of 7 MS signals gave the lowest number of misclassifications, while 7 MS signals (8562.3, 8684.4, 8765.1, 9423.5, 13761.5, 14145.2, and 17250.8 m/z) were extracted as the most shared ones. Using this proteomic model, plasma samples from both PDAC patients and healthy subjects were classified as either positive or negative for cancer, which showed that the sensitivity for prediction was 76.3% (61 of 80 of the cancer patients, 95% CI 65.4–85.1) and for specificity was 91.3% (73 of 80 of the healthy subjects, 95% CI 82.8–96.4, Table 1), for an overall

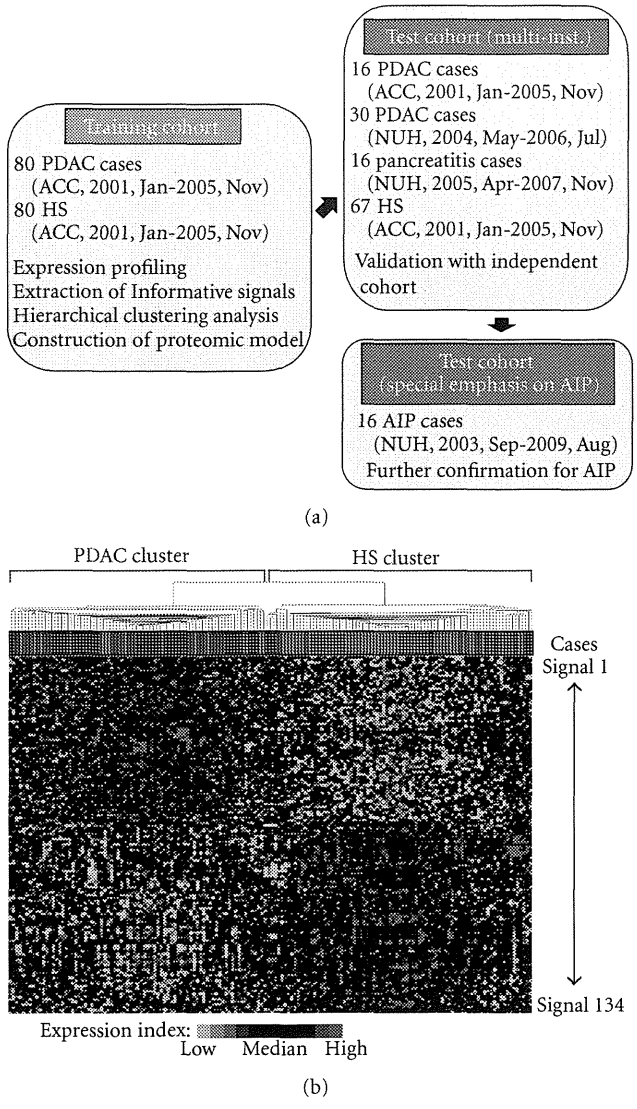

FIGURE 1: MALDI MS analysis of plasma specimens from human PDAC patients and healthy subjects in the training cohort. (a) Independent training-validation-confirmation datasets of 160 training cases, 129 validation cases, and 16 confirmation cases. (b) Unsupervised hierarchical clustering analysis of 80 human PDAC patients and 80 healthy subjects in the training cohort according to the protein expression patterns of 134 MS signals. Each row represents an individual proteomic signal and each column an individual sample. The dendrogram at the top shows the similarities in protein expression profiles among the samples. Substantially elevated (red) expression of the proteins was observed in individual plasma samples. HS: healthy subjects; PDAC: pancreatic ductal adenocarcinoma. Red box case: PDAC: blue box case: healthy subject.

classification accuracy of 83.8% (134 of 160, 95% CI 77.1–89.1). We also calculated positive and negative predictive values (PPV and NPV, resp.) to confirm the diagnostic power of our model, which were 89.8% and 79.3%, respectively. We observed no significant difference for detection of PDAC patients related to lymph node positivity and prognosis. Furthermore, we analyzed the relationship between the age of PDAC patients (≤60 or >60 years old) and detection power of the 7 MS signals. Those results showed that the sensitivity for prediction was 69.8% (30 of 43, 95% CI 53.9–82.8) and 83.8% (31 of 37, 95% CI 68.0–93.8) in the younger and older groups, respectively (Table 1), with no significance in discrimination found ($P = 0.142$, Fisher's exact test). Representative spectra that comprised the 7-signal proteomic model for the healthy subjects and PDAC patients are shown in Figure 2(b). It is of note that our model was able to correctly distinguish 72.7% (8 of 11 cases, 95% CI 39.0–94.0) of the stage I and II cases from the healthy subjects, while it also correctly classified 78.8% (26 of 33, 95% CI 61.1–91.0) of the PDAC patients eligible for surgical resection as positive for cancer (Table 1).

3.2. Protein Expression Profiling in the Test Cohort. It has been well reported that the robustness, including accuracy, of a prediction model should be assessed using an independent validation cohort, even when cross-validation methods, such as LOOCV or n-fold CV, were properly used for developing the prediction model [19]. To examine the robustness of the 7-signal proteomic model constructed with data from MALDI-MS analysis of the training cohort, we applied it to an independent test dataset obtained from plasma samples collected at two different institutions. We also determined whether the identified proteomic model could discriminate between acute and chronic pancreatitis patients, as well as autoimmune pancreatitis, as the discovery of biomarkers applicable for differential diagnosis between PDAC and noncancerous pancreatic diseases has great potential for clinical practice. For the test cohort, plasma samples were obtained from 46 PDAC patients (16 and 30 cases of ACC and NUH, resp.) and 67 healthy subjects from the ACC group, while 16 pancreatitis samples obtained from Nagoya University hospital (NUH) consisted of 2 acute pancreatitis, 11 chronic pancreatitis, and 3 autoimmune pancreatitis cases (Figure 1(a), Supplementary Tables S1 and S2 for additional clinical information for AIP patients). With the 7-signal proteomic model, 82.6% (38 of 46, 95% CI 68.6–92.2) of the cancer cases were classified into the positive group, while 89.2% (74 of 83, 95% CI 80.4–94.9) of the noncancerous subjects were assigned to the group negative for cancer (Figure 3 and Table 2). We calculated PPV and NPV, which were 80.9% and 90.2%, respectively, and the overall accuracy of the classification with the test cohort was 86.8% (112 of 129, 95% CI 79.7–92.1). We also evaluated the relationship between blood vessel invasion (surgery with or without mesenteric venous tract resection) and detection power of the 7 MS signals. Our results showed that the sensitivity for prediction was 88.8% (8 of 9, 95% CI 51.8–99.7) for PDAC patients who underwent mesenteric venous tract resection and 78.6% (11 of 14, 95% CI 49.2–95.3) for those who did not, with no significant difference found ($P = 0.524$, Fisher's exact test). Future studies with a larger number of PDAC patients treated with surgery are warranted to validate the clinical usefulness of our 7-signal proteomic signature. It is of note that our model was able to correctly distinguish 62.5%

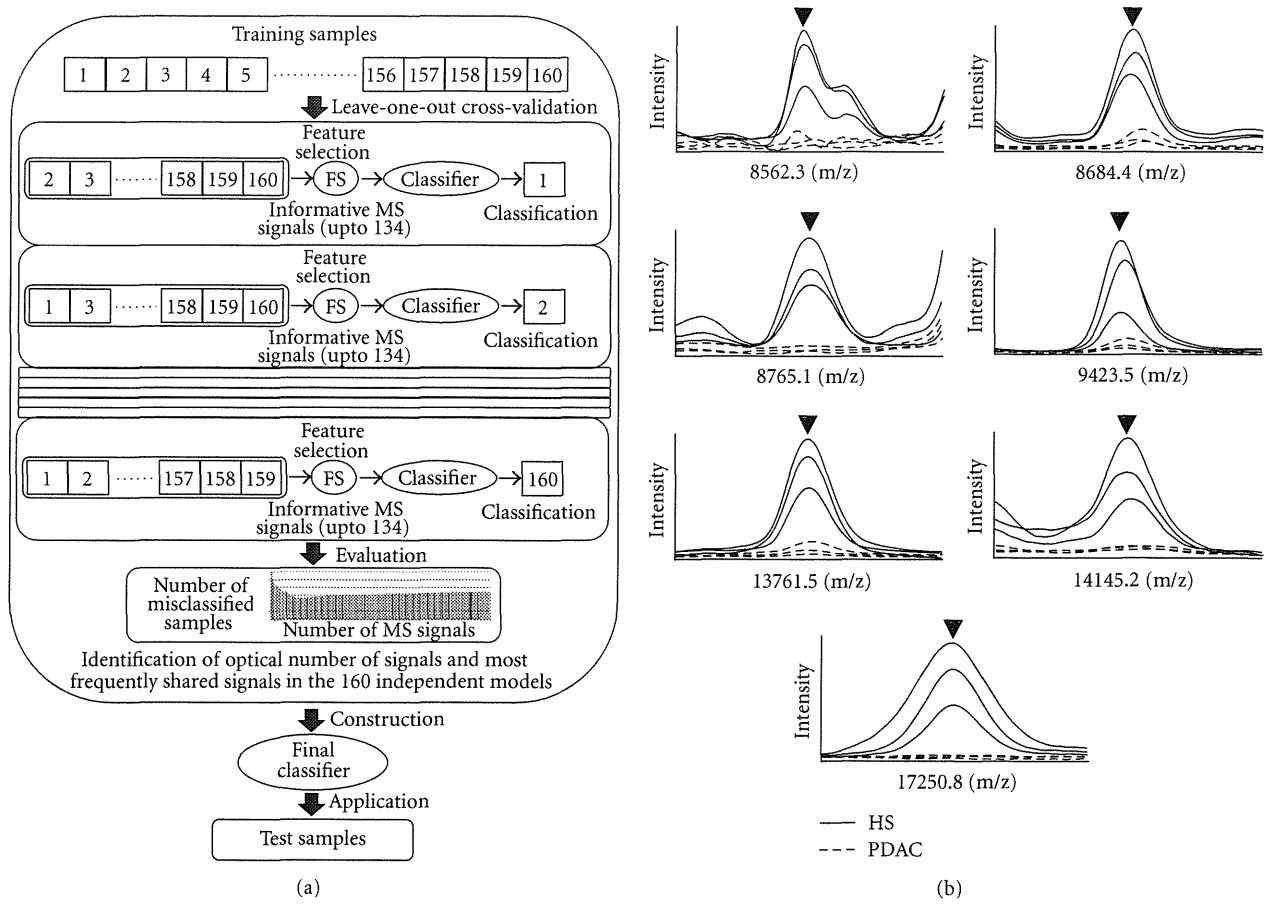

FIGURE 2: Construction of proteomic model for discrimination of PDAC cases from healthy subjects. (a) Schematic diagram of construction of proteomic discrimination model. (b) Representative mass spectra comprising 7-signal proteomic signature. Arrowheads show informative peaks for discrimination between healthy subjects and PDAC patients. Blue lines show representative spectra from healthy subjects and red lines show representative spectra from PDAC patients.

TABLE 1: Discrimination of samples in the training cohort according to 7-signal proteomic model.

	Number of cases analyzed	Number of correctly assigned cases (%)	95% C.I.* (%)
All samples	160	134 (83.8)	77.1–89.1
Pancreatic ductal adenocarcinoma	80	61 (76.3)	65.4–85.1
Healthy subjects	80	73 (91.3)	82.8–96.4
age			
≤60	43	30 (69.8)	53.9–82.8
>60	37	31 (83.8)	68.0–93.8
Clinical stage of pancreatic ductal adenocarcinoma patients			
0/I	3	3 (100)	29.2–100
II	8	5 (62.5)	24.5–91.5
III	8	8 (100)	63.1–100
IVa	14	10 (71.4)	41.9–91.6
IVb	47	35 (74.5)	59.7–86.1

* 95% confidence interval.

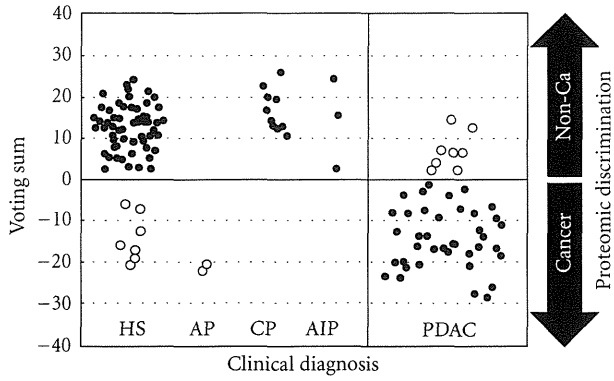

FIGURE 3: Assessment of 7-signal proteomic model with the validation cohort using weighted voting algorithm. The results of proteomic analyses of the training cohort are shown. Each circle represents a voting sum for a single patient. Solid circles: specimens whose prediction with proteomic model matched clinical diagnosis; open circles: specimens whose prediction with proteomic model did not match clinical diagnosis; HS: healthy subjects; AP: acute pancreatitis; CP: chronic pancreatitis; AIP: autoimmune pancreatitis; PDAC: pancreatic ductal adenocarcinoma.

(5 of 8 cases, 95% CI 24.5–91.5) of the stage I and II cases from the healthy subjects and also classified 78.9% (30 of 38, 95% CI 62.7–90.5) of the PDAC patients eligible for surgical resection as positive for cancer. It is also noteworthy that the identified proteomic model distinguished 100% of the patients with chronic pancreatitis (11 of 11, 95% CI 71.5–100) and AIP (3 of 3, 95% CI 29.2−100) from cancer cases (Figure 3 and Table 2).

3.3. Discrimination of Autoimmune Pancreatitis from PDAC Using 7-Signal Proteomic Model. Autoimmune pancreatitis is a systemic inflammatory disease of the pancreas and several diagnostic criteria have been proposed. However, their usefulness is under debate and accurate differential diagnosis remains difficult. Moreover, an important step in diagnosing AIP is to discriminate it from PDAC. In the present study, all (3 of 3) of the AIP patients were correctly discriminated from those with PDAC in the analysis with the test dataset; thus we next performed a confirmatory analysis using plasma samples collected from 16 AIP patients treated at NUH (Figure 1(a) and Supplementary Table S2). For this, we employed our 7-signal proteomic model to investigate whether it would classify the AIP patients as noncancerous and found that it correctly assigned those patients as negative for cancer with 100% accuracy (16 of 16 cases, 95% CI 79.4–100). Therefore, the high potential for discrimination of AIP from PDAC was validated with an independent confirmatory dataset used in a blinded manner. The serum level of CA19-9 was elevated in 4 (21.1%, 95% CI 7.3−52.4) of the AIP cases in our cohort, while IgG4 levels have been reported to be elevated in 10–30% of PDAC cases [7, 20]. Thus, our proteomic model may be applicable as a novel serological test to discriminate AIP from PDAC in clinical practice. Representative spectra obtained from the AIP and PDAC cases are shown in Figure 4.

3.4. Combination of MALDI Proteomic Signature and CA19-9 for Cancer Screening. Our 7-signal proteomic model was able to detect 82.6% (38 of 46, 95% CI 68.6–92.2) of the PDAC patients in the test cohort (Table 2). Moreover, it assigned 78.9% (30 of 38, 95% CI 62.7–90.5) of the patients eligible for an operation to the cancerous group, while 62.5% (5 of 8 and 95% CI 24.5–91.5) of the stage I and II cases were also detected with the identified model. Since it is possible that our 7-signal proteomic model and CA19-9 level are complementary, we investigated whether their combined use would improve the detection rate of patients who may benefit from surgery. The overall sensitivity of CA19-9 (cutoff value, 37 units/mL) alone for stage 0–IVa patients was 71.1% (27 of 38, 95% CI 54.1–84.6), while a combination of our 7-signal proteomic model and CA19-9 level detected 89.5% (34 of 38, 95% CI 75.2–97.1) of operable cases. Notably, for detection of stage I and II PDAC patients, CA19-9 assigned only 50.0% (4 of 8, 95% CI 15.7–84.3) of the cases to the positive group and no additional discrimination power of that marker was observed when combined with our proteomic model. Accordingly, we consider that our 7-signal proteomic model might be more sensitive for detection of early stage PDAC patients than CA19-9, which would improve clinical outcomes following surgical treatment.

3.5. Identification of Individual Proteins in the Proteomic Signature. As an initial step toward elucidating the biologic mechanism of the association between the proteomic signature and carcinogenesis, we identified a couple of proteins that correspond to the mass spectrometry signals in the proteomic signature obtained from serum. Extracts from two serum samples of healthy individual were fractionated by reverse phase-HPLC and analyzed by MALDI MS to identify the HPLC fractions that contained proteins corresponding to peaks in the proteomic signature. These selected fractions were subjected to sequence analysis of tryptic peptides by use of MALDI MS. Accordingly, we identified the following proteins as part of the proteomic signature: apolipoprotein A-I ($[M + H]^+$ = 17,250.8 m/z) and C-III ($[M + H]^+$ = 8765.1), and transthyretin ($[M + H]^+$ = 13761.5).

4. Discussion

In the present study, we analyzed the protein expression profiles of plasma specimens obtained from patients with PDAC, as well as acute and chronic pancreatitis cases, and autoimmune pancreatitis (AIP) patients with MALDI MS. Using bioinformatic analysis, we derived 7 MS signals that allowed us to produce a proteomic model for discrimination of PDAC from noncancerous individuals. When we used our proteomic model with both independent test cohort and confirmation group, 62.5% (5 of 8, 95% CI 24.5–91.5) of stage 0–II cases were correctly assigned to the cancerous group, while all AIP patients (19 of 19, 95% CI 82.4–100)

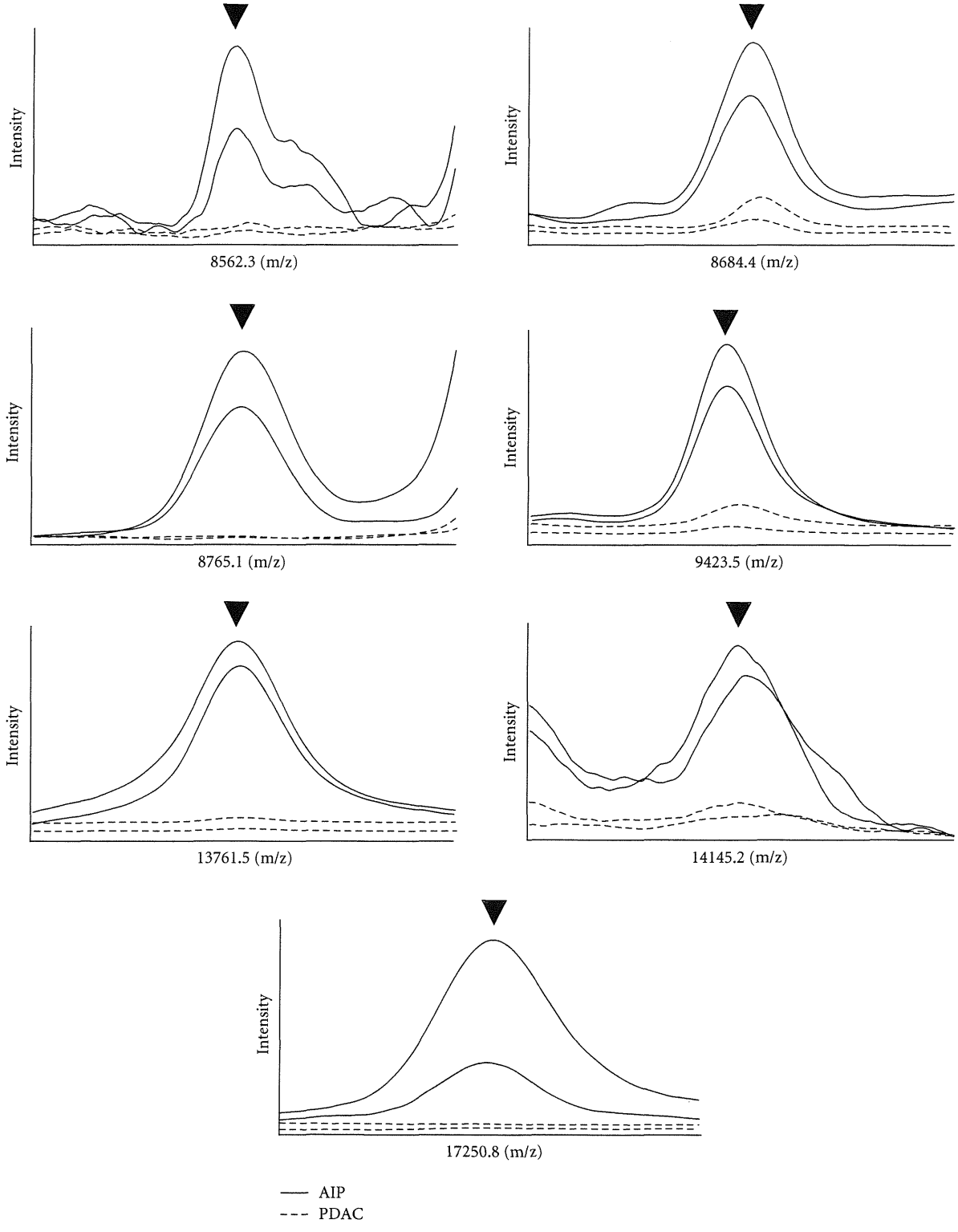

FIGURE 4: Representative mass spectra comprising 7-signal proteomic signature in autoimmune pancreatitis patients and PDAC patients. Arrowheads show informative peaks for discrimination between autoimmune pancreatitis patients and patients with pancreatic cancer. Blue solid and dotted lines show representative spectra from autoimmune pancreatitis patients, and red solid and dotted lines show representative spectra from pancreatic cancer patients. AIP: autoimmune pancreatitis; PDAC: pancreatic ductal adenocarcinoma.

TABLE 2: Discrimination of samples in the test cohort according to 7-signal proteomic model.

	Number of cases analyzed	Number of correctly assigned cases (%)	95% C.I.* (%)
All samples	129	112 (86.8)	79.7–92.1
Healthy subjects	67	60 (89.6)	79.7–95.7
Pancreatic ductal adenocarcinoma (ACCH)	16	13 (81.3)	54.4–96.0
Pancreatic ductal adenocarcinoma (NUH)	30	25 (83.3)	65.3–94.4
Acute pancreatitis (NUH)	2	0 (0)	0–84.2
Chronic pancreatitis (NUH)	11	11 (100)	71.5–100
Autoimmune pancreatitis (NUH)	3	3 (100)	29.2–100
Clinical stage of pancreatic ductal adenocarcinoma patients at ACCH			
0/I	0	NA	NA
II	1	0 (0)	0–97.5
III	3	3 (100)	29.2–100
IVa	4	2 (50)	6.8–93.2
IVb	8	8 (100)	63.1–100
Clinical stage of pancreatic ductal adenocarcinoma patients at NUH			
0/I	1	0 (0)	0–97.5
II	6	5 (83.3)	35.9–99.6
III	13	11 (84.6)	54.6–98.1
IVa	10	9 (90)	55.5–99.7
IVb	0	NA	NA

* 95% confidence interval
NA: not available.

were correctly assigned to the noncancerous group. Discrimination of AIP from cancer is obviously important; however it is currently problematic in clinical practice. Although previous reports have shown discrimination power of proteomic signature between PDAC patients and control subjects [21–24], to the best of our knowledge, the present 7-signal proteomic model is the first system of proteomic prediction based upon mass spectrometry found capable to both detect early-stage PDAC cases and discriminate AIP patients.

Early detection is essential for improving the outcomes of PDAC patients. However, those in stages 0–II are difficult to detect with current diagnostic approaches, including computerized tomography scanning, positron emission tomography scanning, and tissue-based diagnostic tests. CA19-9 is a tumor marker widely used for evaluations of therapeutic effects and detection of PDAC recurrence, though it is not considered to be applicable for mass screening when used alone [4, 6, 25, 26]. Recent advances in molecular biology have also revealed that clinical features cannot be adequately characterized or predicted by a single marker. Thus, microarray analysis has been employed to simultaneously investigate the expression levels of thousands of genes and identify mRNA patterns associated with various human diseases including PDAC [27–29]. However, mRNA expression does not always indicate which of the corresponding proteins are expressed or provide information regarding their post-translational regulation. Moreover, blood and body fluids, such as pancreatic juice and urine, do not contain mRNA. Thus, proteome analysis of such specimens is considered to better reflect the underlying clinical characteristics of human diseases as compared to gene expression profiling, while proteomic technologies including MS have been employed to analyze proteomes in clinical specimens [10–14, 30–32]. Previous proteomics studies of PDAC with healthy controls have shown promising results in distinguishing PDAC, with a sensitivity ranging from 78 to 91% and specificity from 75 to 100% [21–24, 33, 34]. These discrimination power results are better than those obtained with the current CA19-9 marker, while improved diagnostic performance has been observed when serum MS markers were combined with CA19-9 [21, 22, 24]. In the present study, we found that the combination of our 7-signal proteomic model and CA19-9 level improved the positive rate of detection of PDAC patients eligible for surgical resection to 89.5% (34 of 38, 95% CI 75.2–97.1). It is noteworthy that detection of stage I-II cases was also attainable at a sensitivity of 62.5% (5 of 8, 95% CI 24.591.5) without further improvement by adding CA19-9. These results support the usefulness of our 7-signal proteomic model for detection of early stage cases. Since we constructed the present 7-signal model independent from CA19-9, further optimization of selection of a proteomic signature with focus on early detection possibly along with adjustment of the CA19-9 cutoff value is warranted to obtain increased sensitivity. The present 7-signal proteomic model showed high potential to assign inflammatory pancreatic disease patients to the noncancerous group (93.8%; 30 of 32, 95% CI 79.2–99.2). Interestingly, 2 of the misclassified

patients suffered from acute pancreatitis; however, all of the patients of chronic pancreatitis and AIP (11 of 11, 95% CI 71.5–100; and 19 of 19, 95% CI 82.4–100) were correctly assigned to the noncancerous group by our proteomic model. Discrimination of AIP from PDAC is difficult in clinical practice, as symptoms such as obstructive jaundice or space occupying lesions in the pancreas are commonly observed in both cases. Actually, most of the AIP patients in this study showed at least one of these symptoms. Our proteomic model distinguished between AIP patients and those with PDAC with high accuracy; thus it is considered to be effective in future clinical applications, especially for selecting those who are eligible for invasive diagnostic procedures followed by inevitably invasive surgical treatment for PDAC. During the course of our study, Frulloni et al. reported that autoantigens against the plasminogen binding protein of helicobacter pylori and ubiquitin-protein ligase E3 component n-recognin 2 were detected in most of the AIP patients tested, as well as a small number of PDAC cases [35]. It would be interesting to combine our proteomic model with testing for those autoantigens for diagnosis of chronic pancreatic diseases.

In this study, 2 acute pancreatitis patients and 14 healthy subjects were assigned to the cancerous group by our 7-signal proteomic model in the training (7 healthy subjects) and test (2 acute pancreatitis patients and 7 healthy subjects) cohorts. Since that time, we have carefully followed their clinical courses of these healthy subjects and found that 5 suffered from cancerous disease within 3 years, including 2 with rectal cancer, 1 with prostate cancer, 1 with hepatocellular carcinoma, and 1 with a metastatic bone tumor from an unknown primary site. In addition, another false positive healthy subject later developed polyposis in the colon. These observations suggest potential relation of our proteomic model with these malignancies, although further in-depth investigations are apparently required to draw definitive conclusions.

Mass spectrometry profiles obtained from complex protein mixtures can contain thousands of data points derived from real protein signatures. However, they can also be contaminated by electronic and chemical noise, variability in instrumentation, and variable crystallization of the matrix, necessitating careful analytical techniques [11, 13, 14]. In the present study, we employed multiple statistical methods and leave-one-out cross-validation to combine differentially expressed proteins with the clinical variables and found that a minimal set of 7 low-molecular weight proteins was sufficient to distinguish between healthy subjects and PDAC patients. The discriminating power of the extracted proteomic signature was further validated using independent test datasets obtained from plasma specimens collected at 2 different institutions. With this protocol, we carefully eliminated accidental identification of overly optimistic and nonbiological/mathematical multivariate signatures within a closed cohort by overfitting.

The primary goal of this study was development of a bioassay applicable to clinical practice for detection of PDAC and discrimination from AIP, as attempts to identify proteins that comprise a proteomic model have not been fully successful to date. However, the high reproducibility of MALDI MS indicates that direct application of its findings would be successful. In the previous study, Koomen et al. reported that a set of 4 peaks could be used to detect PDAC, of which one MS signal was downregulated in PDAC patients and found to be derived from apolipoprotein A-I [23], while Yan et al. found that transthyretin levels were independently associated with PDAC likelihood when obstructive jaundice was considered [36]. Accordingly, our identification of apolipoprotein A-I and transthyretin, which is a constituent of our proteomic model and downregulated in PDAC patients in this study, is in accord with previous reports from different institutes. We also identified the downregulation of apolipoprotein C-III in serum samples obtained from PDAC patients [37, 38]. Further investigations are warranted to identify discriminating proteins for ascertainment of their functional significance. Notably, 2 downregulated peaks (8765 and 13762 m/z), which were previously extracted as proteomic serum markers for lung cancer [39], were also identified as downregulated proteomic signals in PDAC patients in the present study.

Prospective multi-institutional studies with a larger number of patients including those with early-stage PDAC, AIP, and other pancreatic diseases are apparently warranted to validate further significance of our 7-signal proteomic signature for clinical application. Given that it has potential for early detection of PDAC as well as accurate discrimination of AIP, our 7-signal proteomic model may ultimately lead to a reduction in the large number of deaths caused by devastating cancer and also provide better management for chronic inflammatory disease of pancreas.

Ethical Approval

Requisite approval from the institutional review boards and written informed consent from all subjects were obtained.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Exploratory Research and Program for Improvement of Research Environment for Young Researchers from Special Coordination Funds for Promoting Science and Technology commissioned and a Grant-in-Aid for Scientific Research on Priority Areas, a Grant-in-Aid for Scientific Research (C) by the Ministry of Education, Culture, Sports, Science and Technology of Japan. The sponsors of the study had no role in its design, data collection, analysis, and interpretation of data, the decision to submit the manuscript for publication, or writing the manuscript.

References

[1] http://ganjoho.ncc.go.jp/public/statistics/backnumber/2009_en.html
[2] http://www.cancer.gov/cancertopics/types/pancreatic.
[3] M. Yamamoto, O. Ohashi, and Y. Saitoh, "Japan pancreatic cancer registry: current status," *Pancreas*, vol. 16, no. 3, pp. 238–242, 1998.

[4] M. Goggins, M. Canto, and R. Hruban, "Can we screen high-risk individuals to detect early pancreatic carcinoma?" *World Journal of Surgical Oncology*, vol. 74, no. 4, pp. 243–248, 2000.

[5] R. A. Abrams, L. B. Grochow, A. Chakravarthy et al., "Intensified adjuvant therapy for pancreatic and periampullary adenocarcinoma: survival results and observations regarding patterns of failure, radiotherapy dose and CA 19-9 levels," *International Journal of Radiation Oncology Biology Physics*, vol. 44, no. 5, pp. 1039–1046, 1999.

[6] R. E. Ritts and H. A. Pitt, "CA 19-9 in pancreatic cancer," *Surgical Oncology Clinics of North America*, vol. 7, no. 1, pp. 93–101, 1998.

[7] H. Hamano, S. Kawa, A. Horiuchi et al., "High serum IgG4 concentrations in patients with sclerosing pancreatitis," *The New England Journal of Medicine*, vol. 344, no. 10, pp. 732–738, 2001.

[8] D. L. Finkelberg, D. Sahani, V. Deshpande, and W. R. Brugge, "Autoimmune pancreatitis," *The New England Journal of Medicine*, vol. 355, no. 25, pp. 2670–2676, 2006.

[9] T. Pickartz, J. Mayerle, and M. M. Lerch, "Autoimmune pancreatitis," *Nature Clinical Practice Gastroenterology & Hepatology*, vol. 4, no. 6, pp. 314–323, 2007.

[10] F. Taguchi, B. Solomon, V. Gregorc 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," *Journal of the National Cancer Institute*, vol. 99, no. 11, pp. 838–846, 2007.

[11] E. F. Petricoin, A. M. Ardekani, B. A. Hitt et al., "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572–577, 2002.

[12] B. L. Adam, Y. Qu, J. W. Davis et al., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, vol. 62, no. 13, pp. 3609–3614, 2002.

[13] K. Yanagisawa, Y. Shyr, B. J. Xu et al., "Proteomic patterns of tumour subsets in non-small-cell lung cancer," *The Lancet*, vol. 362, no. 9382, pp. 433–439, 2003.

[14] K. Yanagisawa, S. Tomida, Y. Shimada, Y. Yatabe, T. Mitsudomi, and T. Takahashi, "A 25-signal proteomic signature and outcome for patients with resected non-small-cell lung cancer," *Journal of the National Cancer Institute*, vol. 99, no. 11, pp. 858–867, 2007.

[15] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[16] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[17] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane, "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification," *Journal of the National Cancer Institute*, vol. 95, no. 1, pp. 14–18, 2003.

[18] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[19] A. Dupuy and R. M. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *Journal of the National Cancer Institute*, vol. 99, no. 2, pp. 147–157, 2007.

[20] A. Ghazale, S. T. Chari, T. C. Smyrk et al., "Value of serum IgG4 in the diagnosis of autoimmune pancreatitis and in distinguishing it from pancreatic cancer," *American Journal of Gastroenterology*, vol. 102, no. 8, pp. 1646–1653, 2007.

[21] G. M. Fiedler, A. B. Leichtle, J. Kase et al., "Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer," *Clinical Cancer Research*, vol. 15, no. 11, pp. 3812–3819, 2009.

[22] K. Honda, Y. Hayashida, T. Umaki et al., "Possible detection of pancreatic cancer by plasma protein profiling," *Cancer Research*, vol. 65, no. 22, pp. 10613–10622, 2005.

[23] J. M. Koomen, L. N. Shih, K. R. Coombes et al., "Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins," *Clinical Cancer Research*, vol. 11, no. 3, pp. 1110–1118, 2005.

[24] J. Koopmann, Z. Zhang, N. White et al., "Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry," *Clinical Cancer Research*, vol. 10, no. 3, pp. 860–868, 2004.

[25] F. Safi, W. Schlosser, G. Kolb, and H. G. Beger, "Diagnostic value of CA 19-9 in patients with pancreatic cancer and non-specific gastrointestinal symptoms," *Journal of Gastrointestinal Surgery*, vol. 1, no. 2, pp. 106–112, 1997.

[26] H. Narimatsu, H. Iwasaki, F. Nakayama et al., "Lewis and secretor gene dosages affect CA19-9 and DU-PAN-2 serum levels in normal individuals and colorectal cancer patients," *Cancer Research*, vol. 58, no. 3, pp. 512–518, 1998.

[27] C. A. Iacobuzio-Donahue, A. Maitra, M. Olsen et al., "Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays," *American Journal of Pathology*, vol. 162, no. 4, pp. 1151–1162, 2003.

[28] H. Han, D. J. Bearss, L. W. Browne, R. Calaluce, R. B. Nagle, and D. D. Von Hoff, "Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray," *Cancer Research*, vol. 62, no. 10, pp. 2890–2896, 2002.

[29] B. Ryu, J. Jones, N. J. Blades et al., "Relationships and differentially expressed genes among pancreatic cancers examined by large-scale serial analysis of gene expression," *Cancer Research*, vol. 62, no. 3, pp. 819–826, 2002.

[30] R. M. Caprioli, T. B. Farmer, and J. Gile, "Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS," *Analytical Chemistry*, vol. 69, no. 23, pp. 4751–4760, 1997.

[31] S. A. Schwartz, R. J. Weil, R. C. Thompson et al., "Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry," *Cancer Research*, vol. 65, no. 17, pp. 7674–7681, 2005.

[32] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli, "Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues," *Nature Medicine*, vol. 7, no. 4, pp. 493–496, 2001.

[33] M. Ehmann, K. Felix, D. Hartmann et al., "Identification of potential markers for the detection of pancreatic cancer through comparative serum protein expression profiling," *Pancreas*, vol. 34, no. 2, pp. 205–214, 2007.

[34] J. Guo, W. Wang, P. Liao et al., "Identification of serum biomarkers for pancreatic adenocarcinoma by proteomic analysis," *Cancer Science*, vol. 100, no. 12, pp. 2292–2301, 2009.

[35] L. Frulloni, C. Lunardi, R. Simone et al., "Identification of a novel antibody associated with autoimmune pancreatitis," *The New England Journal of Medicine*, vol. 361, no. 22, pp. 2135–2142, 2009.

[36] L. Yan, S. Tonack, R. Smith et al., "Confounding effect of obstructive jaundice in the interpretation of proteomic plasma profiling data for pancreatic cancer," *Journal of Proteome Research*, vol. 8, no. 1, pp. 142–148, 2009.

[37] H. L. Huang, T. Stasyk, S. Morandell et al., "Biomarker discovery in breast cancer serum using 2-D differential gel electrophoresis/MALDI-TOF/TOF and data validation by routine clinical assays," *Electrophoresis*, vol. 27, no. 8, pp. 1641–1650, 2006.

[38] R. D. Oleschuk, M. E. McComb, A. Chow et al., "Characterization of plasma proteins adsorbed onto biomaterials by MALDI- TOFMS," *Biomaterials*, vol. 21, no. 16, pp. 1701–1710, 2000.

[39] P. B. Yildiz, Y. Shyr, J. S. M. Rahman et al., "Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer," *Journal of Thoracic Oncology*, vol. 2, no. 10, pp. 893–901, 2007.

Oncogene (2012), 1–10

www.nature.com/onc

ONCOGENOMICS

The DNA methylation landscape of small cell lung cancer suggests a differentiation defect of neuroendocrine cells

S Kalari[1,4], M Jung[1,4], KH Kernstine[2,5], T Takahashi[3] and GP Pfeifer[1]

Small cell lung cancer (SCLC) is a disease characterized by aggressive clinical behavior and lack of effective therapy. Owing to its tendency for early dissemination, only a third of patients have limited-stage disease at the time of diagnosis. SCLC is thought to derive from pulmonary neuroendocrine cells. Although several molecular abnormalities in SCLC have been described, there are relatively few studies on epigenetic alterations in this type of tumor. Here, we have used methylation profiling with the methylated-CpG island recovery assay in combination with microarrays and conducted the first genome-scale analysis of methylation changes that occur in primary SCLC and SCLC cell lines. Among the hundreds of tumor-specifically methylated genes discovered, we identified 73 gene targets that are methylated in >77% of primary SCLC tumors, most of which have never been linked to aberrant methylation in tumors. These methylated targets have potential for biomarker development for early detection and therapeutic management of SCLC. SCLC cell lines had a greater number of hypermethylated genes than primary tumors. Gene ontology analysis indicated a significant enrichment of methylated genes functioning as transcription factors and in processes of neuronal differentiation. Motif analysis of tumor-specific methylated regions identified enrichment of binding sites for several neural cell fate-specifying transcription factors including NEUROD1, HAND1, ZNF423 and REST. We hypothesize that two potential mechanisms, loss of cell fate-determining transcription factors by methylation of their promoters and functional inactivation of their corresponding genomic-binding sites by DNA methylation, can promote a differentiation defect of neuroendocrine cells thus enhancing the ability of tumor progenitor cells to transition toward SCLC.

Oncogene advance online publication, 20 August 2012; doi:10.1038/onc.2012.362

Keywords: DNA methylation; small cell lung cancer; differentiation; epigenetics

INTRODUCTION

Lung cancer is divided by histology into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). SCLC represents about 15% of all lung cancer cases and is one of the most lethal forms of cancer with properties of high mitotic rate and early metastasis.[1] It is distinctly characterized by small cells with poorly defined cell borders and minimal cytoplasm, rare nucleoli and finely granular chromatin. Although SCLC patients initially respond to chemotherapy and radiation therapy, the disease recurs in the majority of patients. Because of the aggressiveness of SCLC and the lack of effective therapy and early diagnosis, without treatment the median survival time for SCLC is only 2–4 months. With current treatment modalities, the median survival times for limited-stage disease, <5% of the total, is 16–24 months and for extensive disease, 7–12 months, in spite of the fact that 60–80% of patients respond to therapy. It is essential to gain a better understanding of the molecular pathogenesis of the disease and to identify molecular alterations, which could lead to improved results in early detection and a means of assessing response to therapy.

Several studies have identified abnormalities within tumor suppressor genes, oncogenes, signaling pathways, receptor kinases and growth factors that have a proven role in the pathogenesis of various other human cancers. About 90% of SCLC patients' DNA samples have mutations in the *TP53* gene.[2,3] Similarly, another tumor suppressor gene, retinoblastoma, is either deleted or mutated in the majority (about 90%) of SCLCs.[2,4] In addition, higher expression of the *MYC* family of oncogenes has been found in SCLC cell lines, xenografts and fresh tumor specimens.[5–7] Abnormalities in various receptor tyrosine kinase families are commonly found in the majority of SCLC cases. These changes are associated with a more aggressive tumor growth, resistance to therapy and poor prognosis.[8,9] The phosphoinositide 3-kinase/AKT pathway is defective in SCLC patients' tumors. Nearly two thirds of SCLCs have phosphorylated AKT[9] and this constitutively active kinase can modulate a variety of cellular functions such as cell proliferation, survival, motility, adhesion and differentiation.[8] The cellular origin of SCLC is yet to be proven definitively. Recent studies in mice indicated that neuroendocrine cells seem to be the predominant cells of origin of SCLC.[10,11]

SCLC is also characterized by common deletion of the *fragile histidine triad (FHIT)* gene, located at 3p14.[2] Similarly, chromosome 3p21 is another locus, which is frequently subjected to loss in almost all SCLCs, and this event is thought to be an early event in lung cancer pathogenesis.[12] At 3p21.3, there are several candidate tumor suppresser genes, including the Ras association domain family member 1A (*RASSF1A*), tumor suppressor candidate 2

[1]Department of Cancer Biology, Beckman Research Institute of the City of Hope, Duarte, CA, USA; [2]Department Surgery, Beckman Research Institute of the City of Hope, Duarte, CA, USA and [3]Division of Molecular Carcinogenesis, Nagoya University Graduate School of Medicine, Nagoya, Japan. Correspondence: Dr GP Pfeifer, Department of Cancer Biology, Beckman Research Institute, City of Hope, Duarte, CA 91010, USA.
E-mail: gpfeifer@coh.org
[4]These authors contributed equally to this work.
[5]Current address: Division of Thoracic Surgery, University of Texas Southwestern Medical Center, Dallas, TX, USA.
Received 10 September 2011; revised 18 May 2012; accepted 4 July 2012

(*TUSC2*, also known as *FUS1*), semaphorin 3B (*SEMA3B*) and semaphorin 3F (*SEMA3F*).[13,14]

In contrast to the genetic alterations discussed above, epigenetic aberrations, specifically DNA methylation changes found in SCLC tumors, have not been studied so far in a comprehensive manner. DNA methylation analysis might provide vital information that could shed light on mechanisms of disease initiation, development and progression, as well as lead to cancer biomarker discovery.[15,16] There are several gene-specific DNA methylation studies for SCLC. For example, promoter hypermethylation of the tumor suppressor gene *RASSF1A* and subsequent suppression of its expression is found in almost all of the SCLC tumors.[17,18] Another study found *caveolin-1* (*CAV1*) gene methylation in over 90% the tested SCLC cell lines.[19]

Lack of genome-wide DNA methylation studies in SCLC prompted us to undertake this task. We applied the methylated-CpG island recovery assay (MIRA), which has shown excellent sensitivity for identification of methylated genomic regions in cancer,[20–23] to map DNA methylation patterns at promoters and CpG islands of primary SCLC tumors, SCLC cell lines and normal lung control samples.

RESULTS

Identification of methylated genes in human SCLC tissue on a genome-wide platform

The MIRA technique, used in combination with microarray analysis, is a high-resolution mapping technique and has proven successful for profiling global DNA methylation patterns in NSCLC and other tumors.[22–25] In this study, we have applied this sensitive method to study the methylation status of CpG islands and promoters in SCLC to investigate the potential role of methylation changes in the initiation and development of SCLC, as well as to discover potential biomarkers for better management of the disease. Eighteen human primary SCLC and five SCLC cell line DNA samples were screened for methylation by MIRA-based microarrays. DNAs from five normal healthy lung tissues adjacent to the tumor and obtained at the time of surgical resection were used as controls in the MIRA analysis. DNA was subjected to MIRA enrichment as described previously[26,27] and subsequent microarray analysis was performed on 720k Nimblegen CpG island plus promoter arrays.

Microarray data analysis

To increase the specificity of MIRA-based enrichment signals, we chose to call peaks based on different quantiles of four neighboring probes. Peaks were then calculated using the base functions of the Bioconductor package Ringo.[28] Table 1 shows the specificity and sensitivity of this approach relative to different quantile ranges using DNA from the SCLC cell line SW1271. Based on the validations conducted by combined bisulfite restriction analysis (COBRA) single-gene methylation assays, we chose an 80% cutoff for medium to strongly methylated regions and a cutoff below 56% defined as not methylated. Thus, compared with the conventional NimbleScan method using the default settings, we could increase the sensitivity of methylation peak detection to 94% without decreasing specificity. As this threshold was defined for one SCLC cell line, we tested the same settings for primary small lung cancer samples and did not observe a significant increase of false positive predicted hypermethylated regions.

Using the peak identification algorithm described in the Materials and methods section, we identified ~15 000 methylation peaks in each sample (Supplementary Table 1). Our clustering analysis of tumor samples and controls showed that SCLC cell lines clustered together and that four of the five normal samples were close to each other, but different tumor samples occupied different spaces in the dendrogram (Supplementary Figure 1).

Table 1. Validation of microarray results by COBRA assays

Top quantile (%)	*No. of targets tested*[a]	*Met*	*UnMet*	*PCR fails*	*% Met*	*% UnMet*
99	10	9	—	1	100	0
95	10	9	—	1	100	0
90	10	9	1	—	90	10
85	10	9	1	—	90	10
80	10	7	3	—	70	30
70	14	3	5	6	37.5	62.5
60	19	3	12	4	20	80
50	13	2	11	—	15	85

Abbreviations: COBRA, combined bisulfite restriction analysis; Met, methylated; UnMet, unmethylated. [a]COBRA was performed for each quantile category with bisulfite-converted DNA from the SW1271 cell line. Results were tabulated for number of Met and UnMet genes in these various categories.

Taking into account that we had 18 tumor samples and 5 normal samples for microarray data analysis, we defined a stringent tumor-specific methylated region as the overlapping region that meets the minimum 80% quantile criterion in 14 of 18 tumors and is below the 56% quantile in 4 of 5 normal tissues. A less stringent set was defined as an overlap between at least 6 peaks from tumor samples out of 18, using the same criteria as above. Thus, we were mainly comparing strongly methylated regions versus poorly methylated regions. Although small methylation level differences could not be picked up this way, the aim of discovering uniquely strongly methylated and tumor-specific regions was well supported by this approach.

Methylated genes in primary SCLC

Supplementary Figure 2 shows examples of tumor-specific methylation peaks at the *PROX1, CCDC140, PAX3 and SIM1* genes located on chromosomes 1, 2 and 6, respectively. Supplementary Figure 3 shows extensive tumor-specific methylation of the *HOXD* cluster on chromosome 2. Compilation of tumor-specific methylation peaks revealed a total of 698 regions in 6 out of 18 tumors (≥33% of SCLC tumors) compared with normal lung DNA, which represented 339 ensembl gene IDs for promoter-related tumor-specifically methylated regions (defined as −5000 to +1000 relative to the TSS), 197 ensembl gene IDs related to peaks mapped to the gene bodies and 63 ensembl gene IDs for peaks mapped downstream of the corresponding genes (Figure 1a; Supplementary Table 2). Individual primary SCLCs contained between 366 and almost 1500 tumor-specific methylation peaks (Supplementary Table 3).

There were 73 tumor-specific methylated peaks, which were found in at least 14 out of 18 SCLC tumors (>77% of SCLC tumors), that corresponded to 28 ensembl gene IDs for promoters, 30 ensembl gene IDs for gene bodies and 11 for downstream regions (Figure 1b). These methylated genes from 77% or more of the SCLC tumors are presented in Table 2 and in Supplementary Table 4, for more detailed information.

Identification of methylated genes in human SCLC lines

Owing to the limited availability of primary SCLC tissue, we added several SCLC cell lines originally derived from primary tumor sites. Owing to the unavailability of neuroendocrine cells, which are believed to be the cell of origin of SCLC,[10] we chose normal bronchial epithelial cells as a control for these studies. Clustering analysis based on the total methylation peaks of SCLC cell lines showed that all cell lines cluster tightly together (Supplementary Figure 1). Further analysis of these methylated peaks for tumor cell line-specific peaks revealed 1223 unique tumor-specific peaks found in 4 out of 5 SCLC cell lines (≥80% of SCLC cell lines)

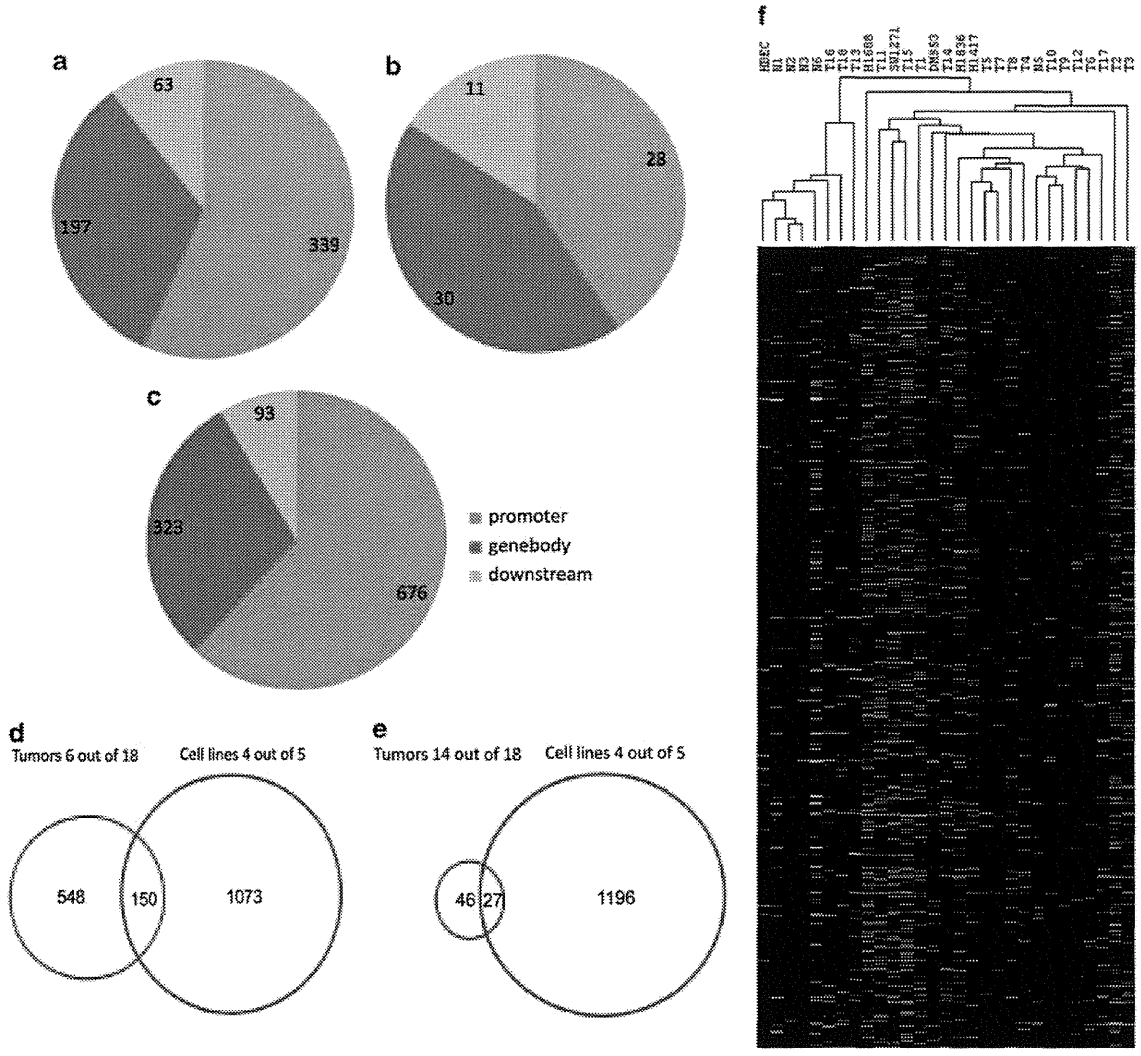

Figure 1. Mapping of tumor-specific methylation peaks in primary SCLC and SCLC cell lines. (**a**) Localization of the methylation peaks in primary SCLC (6 or more out of 18 tumors methylated; that is, peaks that meet the minimum 80% quantile criterion in 6 of 18 tumors) relative to gene position. (**b**) Localization of the methylation peaks in primary SCLC (14 or more out of 18 tumors methylated) relative to gene position. (**c**) Localization of the methylation peaks in SCLC cell lines (4 or more out of 5 cell lines methylated) relative to gene position. (**d**) Overlap of methylation peaks between SCLC primary tumors (6 or more out of 18 tumors methylated) and SCLC cell lines (4 or more out of 5 cell lines methylated). (**e**) Overlap of methylation peaks between SCLC primary tumors (14 or more out of 18 tumors methylated) and SCLC cell lines (4 or more out of 5 cell lines methylated). (**f**) Cluster analysis of methylation peaks. Methylation peaks found in at least 33% of tumor samples but not in normal samples were identified. Then the data were subjected to hierarchical clustering with Euclidean distance and average linkage method using Cluster v3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm) and visualized in Java TreeView (http://jtreeview.sourceforge.net). Red, methylated state; green, unmethylated state.

compared with methylated peaks form normal bronchial epithelial cells (Supplementary Table 5). These peaks represented 676 ensembl gene IDs mapped to promoter regions, 323 ensembl gene IDs corresponding to methylated regions in the gene body and 93 ensembl gene IDs where the hypermethylated regions could be located downstream of genes (Figure 1c). Individual cell lines contained between 2779 and 4485 cell line-specific methylation peaks (Supplementary Table 3), numbers that were greater than those found in primary SCLCs. We compared SCLC tumor-specific methylated regions with SCLC cell line-specific methylated regions. There was a relatively small group (<20%) of SCLC cell line-specific genes found to be commonly (>6 of 18) methylated in primary SCLC tumors and vice versa (that is, ~21% of SCLC primary tumor peaks matched with those of frequent SCLC cell line methylation; Figure 1d). When we determined the overlap between peaks methylated in 14/18 tumors and 4 of 5 cell lines, the number of overlapped genes was 27 (Figure 1e). We mapped the location of tumor-specific methylation peaks relative to promoters, gene bodies and locations downstream of genes (Figures 1a–c). The distribution patterns were similar for peaks found in ⩾6/18 tumors and in cell lines, but for the most frequently methylated genes (⩾14/18) the peaks tended to be more commonly localized in gene bodies and downstream (Figure 1b). Cluster analysis of methylation peaks in normal and tumor samples is shown in Figure 1f.

Validation of gene-specific methylation in SCLC samples

We further validated tumor-specific methylation peaks discovered by microarray analysis for several of the targets by the COBRA assay. In this assay, bisulfite-converted DNA is PCR-amplified using gene-specific primers and is then digested with a restriction

Table 2. Gene targets methylated in 77% or more of primary SCLCs

Chromosome	*Start peak*	*End peak*	*hgnc_symbol*	*Description*
6	27647872	27648246		
1	91189238	91189687	BARHL2	BarH-like homeobox 2 [Source:HGNC Symbol;Acc:954]
10	124901911	124902685	HMX2	H6 family homeobox 2 [Source:HGNC Symbol;Acc:5018]
15	53087134	53087683	ONECUT1	One cut homeobox 1 [Source:HGNC Symbol;Acc:8138]
9	100611180	100611554	FOXE1	Forkhead box E1 (thyroid transcription factor 2) [Source:HGNC Symbol;Acc:3806]
17	59529794	59530268	TBX4	T-box 4 [Source:HGNC Symbol;Acc:11603]
1	214153078	214153777	PROX1	Prospero homeobox 1 [Source:HGNC Symbol;Acc:9459]
14	95239173	95240547	GSC	Goosecoid homeobox [Source:HGNC Symbol;Acc:4612]
21	38068981	38069055	SIM2	Single-minded homolog 2 (Drosophila) [Source:HGNC Symbol;Acc:10883]
6	117584283	117584857	VGLL2	Vestigial-like 2 (Drosophila) [Source:HGNC Symbol;Acc:20232]
14	37124350	37124799	PAX9	Paired box 9 [Source:HGNC Symbol;Acc:8623]
2	177004205	177004604		
14	36991675	36992549	NKX2-1	NK2 homeobox 1 [Source:HGNC Symbol;Acc:11825]
1	197879403	197880252	LHX9	LIM homeobox 9 [Source:HGNC Symbol;Acc:14222]
11	32455050	32455624	WT1-AS	WT1 antisense RNA (non-protein coding) [Source:HGNC Symbol;Acc:18135]
13	112719925	112720174	SOX1	SRY (sex-determining region Y)-box 1 [Source:HGNC Symbol;Acc:11189]
21	38069706	38069780	SIM2	Single-minded homolog 2 (Drosophila) [Source:HGNC Symbol;Acc:10883]
2	176956605	176956754	HOXD13	Homeobox D13 [Source:HGNC Symbol;Acc:5136]
9	129566330	129566704	ZBTB43	Zinc finger and BTB domain containing 43 [Source:HGNC Symbol;Acc:17908]
3	172167182	172167256	GHSR	Growth hormone secretagogue receptor [Source:HGNC Symbol;Acc:4267]
1	230777303	230777452	COG2	Component of oligomeric golgi complex 2 [Source:HGNC Symbol;Acc:6546]
3	27765097	27765996	EOMES	Eomesodermin [Source:HGNC Symbol;Acc:3372]
20	30639265	30639939	HCK	Hemopoietic cell kinase [Source:HGNC Symbol;Acc:4840]
3	183274057	183274331	KLHL6	Kelch-like 6 (Drosophila) [Source:HGNC Symbol;Acc:18653]
12	114846668	114847217	TBX5	T-box 5 [Source:HGNC Symbol;Acc:11604]
4	122685401	122685475		Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:E7ENT1]
2	182547581	182547655		
20	44880344	44880693	CDH22	Cadherin 22, type 2 [Source:HGNC Symbol;Acc:13251]
9	21402751	21403100	IFNA12P	Interferon, alpha 12, pseudogene [Source:HGNC Symbol;Acc:5443]
7	97360940	97362189	TAC1	Tachykinin, precursor 1 [Source:HGNC Symbol;Acc:11517]
2	223162732	223163206	CCDC140	Coiled-coil domain containing 140 [Source:HGNC Symbol;Acc:26514]
7	129422815	129423514		
2	192711381	192711755		
6	27107272	27107346	HIST1H4I	Histone cluster 1, H4i [Source:HGNC Symbol;Acc:4793]
2	176969205	176970504	HOXD11	Homeobox D11 [Source:HGNC Symbol;Acc:5134]
19	9608951	9609250	ZNF560	Zinc finger protein 560 [Source:HGNC Symbol;Acc:26484]
7	27282651	27282900	EVX1	Even-skipped homeobox 1 [Source:HGNC Symbol;Acc:3506]
2	223163332	223163406	PAX3	Paired box 3 [Source:HGNC Symbol;Acc:8617]
7	27282951	27283025	EVX1	Even-skipped homeobox 1 [Source:HGNC Symbol;Acc:3506]
7	8474326	8475225	NXPH1	Neurexophilin 1 [Source:HGNC Symbol;Acc:20693]
4	174452351	174452925		Nbla00301 (NBLA00301), non-coding RNA [Source:RefSeq DNA;Acc:NR_003679]
4	13545178	13545427	NKX3-2	NK3 homeobox 2 [Source:HGNC Symbol;Acc:951]
X	111325120	111325194	TRPC5	Transient receptor potential cation channel, subfamily C, member 5 [Source:HGNC Symbol;Acc:12337]
6	100911555	100911904	SIM1	Single-minded homolog 1 (Drosophila) [Source:HGNC Symbol;Acc:10882]
14	29243250	29243899	C14orf23	Chromosome 14 open-reading frame 23 [Source:HGNC Symbol;Acc:19828]
5	172660770	172660844	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila) [Source:HGNC Symbol;Acc:2488]
2	220196257	220197006	RESP18	Regulated endocrine-specific protein 18 homolog (rat) [Source:HGNC Symbol;Acc:33762]
9	126776030	126776479	LHX2	LIM homeobox 2 [Source:HGNC Symbol;Acc:6594]
1	165323302	165323951	LMX1A	LIM homeobox transcription factor 1, alpha [Source:HGNC Symbol;Acc:6653]
2	119603031	119603180	EN1	Engrailed homeobox 1 [Source:HGNC Symbol;Acc:3342]
12	63543634	63544008	AVPR1A	Arginine vasopressin receptor 1A [Source:HGNC Symbol;Acc:895]
8	97170050	97170499	GDF6	Growth differentiation factor 6 [Source:HGNC Symbol;Acc:4221]
1	47694839	47695213	TAL1	T-cell acute lymphocytic leukemia 1 [Source:HGNC Symbol;Acc:11556]
13	84453425	84453824	SLITRK1	SLIT and NTRK-like family, member 1 [Source:HGNC Symbol;Acc:20297]
4	174448251	174448725	HAND2	Heart and neural crest derivatives expressed 2 [Source:HGNC Symbol;Acc:4808]
2	176977280	176977729	HOXD10	Homeobox D10 [Source:HGNC Symbol;Acc:5133]
5	37835994	37836168	GDNF	Glial cell-derived neurotrophic factor [Source:HGNC Symbol;Acc:4232]
9	37029751	37030525	PAX5	Paired box 5 [Source:HGNC Symbol;Acc:8619]
14	29247325	29247499	C14orf23	Chromosome 14 open reading frame 23 [Source:HGNC Symbol;Acc:19828]
7	8483051	8483825	NXPH1	Neurexophilin 1 [Source:HGNC Symbol;Acc:20693]
6	154360508	154360857	OPRM1	Opioid receptor, mu 1 [Source:HGNC Symbol;Acc:8156]
20	58569381	58569455	CDH26	Cadherin 26 [Source:HGNC Symbol;Acc:15902]
9	21968201	21968875	C9orf53	Chromosome 9 open reading frame 53 [Source:HGNC Symbol;Acc:23831]
16	49311725	49312274	CBLN1	Cerebellin 1 precursor [Source:HGNC Symbol;Acc:1543]
8	9756191	9756540	MIR124-1	MicroRNA 124-1 [Source:HGNC Symbol;Acc:31502]
5	170741921	170741995	TLX3	T-cell leukemia homeobox 3 [Source:HGNC Symbol;Acc:1353]
20	21488326	21488925	NKX2-2	NK2 homeobox 2 [Source:HGNC Symbol;Acc:7835]
5	170743496	170744170	TLX3	T-cell leukemia homeobox 3 [Source:HGNC Symbol;Acc:13532]

Table 2 (Continued)

Chromosome	*Start peak*	*End peak*	*hgnc_symbol*	*Description*
5	172672295	172672844		Y RNA [Source:RFAM;Acc:RF00019]
2	177027180	177027529	HOXD4	Homeobox D4 [Source:HGNC Symbol;Acc:5138]
4	85402627	85403376	NKX6-1	NK6 homeobox 1 [Source:HGNC Symbol;Acc:7839]
15	96911497	96912071	MIR1469	MicroRNA 1469 [Source:HGNC Symbol;Acc:35378]
15	89949372	89949871	MIR9-3	MicroRNA 9-3 [Source:HGNC Symbol;Acc:31646]
3	50377447	50378846	RASSF1A ([a])	Ras association (RalGDS/AF-6) domain family member 1 [Source:HGNC Symbol;Acc:9882]

[a]Indicates a previously validated gene with a lower threshold for normal tissues than used for the other regions.

endonuclease, either *Bst*UI or *Taq*I, which recognize the sequences 5′-CGCG-3′ or 5′TCGA-3′, respectively. The cytosines in unmethylated restriction sites are converted by sodium bisulfite, amplified by PCR and resist digestion, whereas methylated sites remain unchanged and are cleaved by these enzymes. The digested fragments visualized on agarose gels are thus indicative of methylated restriction sites in the region analyzed. We performed extensive validation analysis by COBRA to confirm the tumor-specific methylated regions (Supplementary Figure 4). Representative examples of COBRA results are shown for the genes *DMRTA2, MIR-129-2* and *GALNTL1*. In total, we inspected the methylation status of 11 genes (*GALNTL1, MIR-10A, MIR-129-2, MIR-196A2, MIR-615, MIR-9-3, AMBRA1, HOXD10, PROX1, ZNF672* and *DMRTA2*) based on the various degrees of methylation obtained from the list of differentially methylated targets. Results for all the targets are presented in Supplementary Table 6. The COBRA analysis revealed that our microarray analysis is highly reliable with over 93% accuracy and only ~4% false negative and ~3% false positive hits.

To further confirm the COBRA results of the methylated genes *GALNTL1* and *DMRAT2*, we sequenced bisulfite-converted DNA from SCLC tumor and matched normal lung samples (Supplementary Figure 4). Normal control lung DNA samples showed either no or very low levels of methylation across the CpG dinucleotides tested in contrast to SCLC tumor DNA samples, which were heavily methylated.

Gene expression and methylation status

For the SCLC cell lines SW1271, H1836 and H1688, and HBECs, Affymetrix gene expression analysis was performed and hypermethylated regions in the SCLC cell lines were compared with their associated probe expression changes. On a global level, we could not detect a correlation between the tumor-specific hypermethylated regions and downregulation of associated genes. This phenomenon has been observed in other tumor methylation studies. Some of the reasons for this lack of correlation are that (1) genes that become methylated in tumors frequently are already expressed at very low levels in corresponding normal tissues,[29–32] (2) methylation-independent mechanisms (such as chromatin modifications) are responsible for expression changes[33] and (3) methylation of alternative promoters obscures such correlations.[27,34] Unlike the methylation patterns, the expression signals of the individual tumor cell lines were not highly correlated to each other when compared with the control cell line (as seen by principal component analysis; data not shown).

Functional pathway analysis of methylated genes

For the two stringencies that were defined ($\geqslant$6 out of 18 tumors specifically hypermethylated and $\geqslant$14 out of 18 tumors specifically hypermethylated), we performed a functional annotation clustering, for promoter proximal tumor-specifically methylated regions and gene body-associated tumor-specifically methylated regions. For $\geqslant$6 out of 18 tumor-specific promoter proximal methylated regions, two main annotation clusters could be identified, one for homeobox genes (*P*-value 1.6E − 26, Bonferroni corrected) and one for transcription factors in general (1.0E − 09; Figure 2a; Supplementary Table 7). More specifically, clusters for neuronal fate commitment (1.3E − 5), neuronal differentiation (3.5E − 9) and pattern specification processes (2.3E − 11) showed the strongest enrichment. In comparison, hypermethylated regions in gene bodies showed similar functional enrichment clusters for homeobox genes (6.2E − 26) and pattern specification processes (3.8E − 11), but significantly less enrichment for neuronal fate commitment (7.0E − 1) and for neuronal differentiation (1.2E − 4; Supplementary Table 8), suggesting that the latter functional categories are more related to promoter-specific methylation (Figure 2a).

Concerning functional enrichment for tumor-specifically hypermethylated regions for the majority of tumors ($\geqslant$14 out of 18 tumors), clusters with significantly less enrichment compared with their less significant counterpart ($\geqslant$6 out of 18) could only be obtained for homeobox genes (7.5E − 7 for promoter regions and 2.3E − 8 for gene bodies) and transcription factors (2.8E − 4 for promoter regions and 3.6E − 2 for gene bodies), which can be partly explained by the lower number of genes in this category (Supplementary Tables 9 and 10). Lung development was another significantly enriched category for promoter methylation (Supplementary Table 9).

With regard to the cell lines, genes associated with hypermethylated regions in the five SCLC cell lines compared with the control cell line, homeobox-related functional terms and transcription factor-related terms were significantly enriched only for gene body-associated tumor peaks (4.8E−8 for homeobox genes and 3.0E − 3 for transcription factors, Bonferroni corrected) but the strong enrichment for these categories observed for promoter regions in the tumor tissues was not present for the cell line models (Supplementary Tables 11 and 12). This probably reflects a greater number and higher diversity of methylation events observed in the cell lines.

For targets methylated simultaneously in $\geqslant$14 out of 18 tumors and in $\geqslant$4 out of 5 cell lines (Supplementary Table 13), we again observed an enrichment in the same functional categories. Notably, this group of genes contained a number of genes involved in neuronal or neuroendocrine differentiation, such as EOMES/TBR2, the gene *TAC1*, which encodes the neuropeptide substance P, and *RESP18*, encoding a neuroendocrine-specific protein.

Motif discovery

We next used the *de novo* motif discovery algorithm HOMER[35] to search for sequence patterns that are associated with regions that are specifically methylated in SCLC tumor samples for at least 33% of the tumors and were able to identify a set of nonredundant sequence motifs that were highly enriched in comparison with all non-tumor-specifically methylated regions on the array.

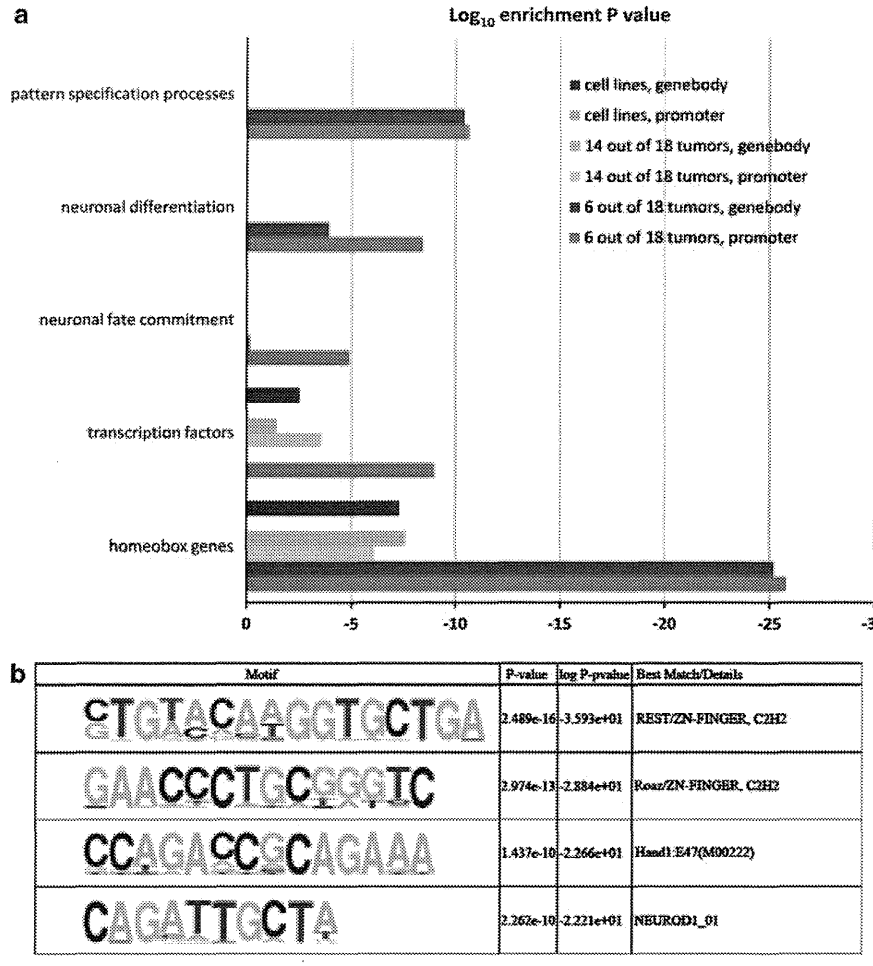

Figure 2. Functional annotation and motif finding analysis. (**a**) Shown are DAVID functional analysis clusters that contained the highest enrichment scores in all three categories: 33% or more of tumors, 77% or more of tumors and cell lines. For more details see Materials and methods. (**b**) Motif finding analysis. Significantly enriched consensus motifs for REST, Roaz/ZNF423, Hand1 and NEUROD1 are shown.

Transcription factors, which were falling into this category, were REST/NRSF (2.5E − 16), ZNF423 (3.0E − 13), HAND1 (1.44E − 10) and NEUROD1 (2.3E − 10; Figure 2b). Examples of methylated NEUROD1 targets are shown in Figure 3. The majority of the sequence motifs identified in methylated regions were enriched within the proximal promoter regions of known genes. The highest enrichment was based on redundant sequence structures and for those that were not, we demanded a stringent alignment with matching transcription factor-binding sites and a low number of occurrences in the background set, which contained all possible methylation sites. REST, ZNF423, HAND1 and NEUROD1 contained nonredundant sequences, a maximal mismatch of 2 bp to the identified *de novo* motif and were selectively enriched in the target sequence set. As such, the identified motifs might not be representative for the whole tumor-specific target set but shed light on sub-regulatory networks with a possibly major impact on the phenotype of SCLC. For example, NEUROD1- and HAND1-binding sites were found in methylated targets representing genes involved in neuronal cell fate commitment such as GDNF, NKX2-2, NKX6-1, EVX1 and SIM2 (Supplementary Tables 2 and 14). Methylation of these binding sites suggests a model in which these transacting factors were lost during tumorigenesis rendering their target sites susceptible to methylation. To analyze this scenario further, we focused on the NEUROD1 transcription factor. Indeed, expression of NEUROD1 proved to be undetectable by a sensitive reverse transcription–PCR assay (Supplementary Figure 5) in the four SCLC cell lines tested and it was expressed at very low levels in human bronchial epithelial cells. In SCLC cell lines and, importantly, also in primary SCLC tumors, the promoter of NEUROD1 was heavily methylated (Supplementary Figures 6A and B) consistent with a possible lack of expression. In addition, we found increased methylation at the promoters of HAND1 and REST in SCLC cell lines and in primary tumors (Supplementary Figure 6).

DISCUSSION

To identify frequently methylated genes in SCLC tumor patients and SCLC cell lines, we have combined the use of a sensitive method for identifying methylation in CpG-rich regions, the MIRA assay[26,27] with genome-wide CpG island and promoter array analysis. Global profiling of 18 SCLC tumor samples compared with normal lung samples resulted in 698 and 73 tumor-specifically methylated and ensembl-annotated gene targets for 33% or more ($\geqslant$6 of 18) of tumors, representing a substantial subgroup of patients, and in 77% or more of SCLC tumors (methylation in at least 14 of 18 samples), representing the majority of all patients, respectively. The 73 gene targets methylated in such a large fraction of the patient population may be of particular value for designing DNA methylation-based biomarkers for early detection of SCLC, for example, in serum or sputum, and for disease management.

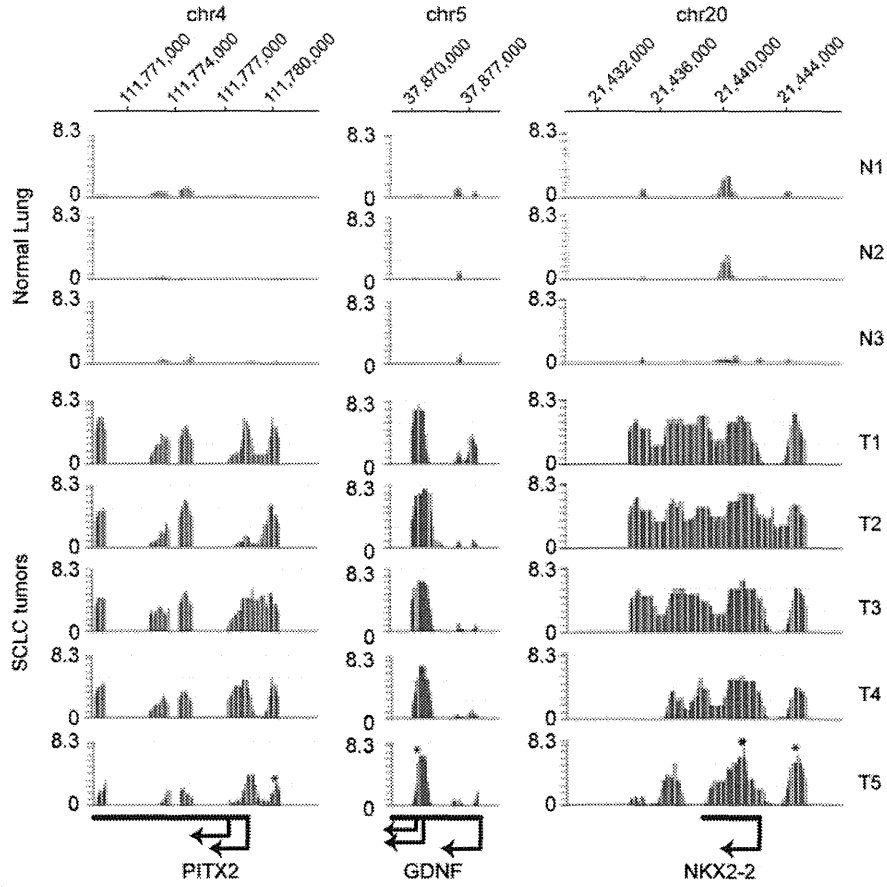

Figure 3. Examples of tumor-specific methylation of NEUROD1 target genes in SCLC. The top of the figure indicates the chromosomal coordinates according to the UCSC Genome browser hg19. Gene names and direction of transcription are shown at the bottom of the figure. The Nimblegen array data (methylated fraction versus input) are shown for three normal lung tissues (red) and five primary SCLC tumors (blue). The methylation signal is shown plotted along the chromosome as a *P*-value score. Therefore, the minimum number on the *y* axis is 0 (when $P = 1$). The *P*-value score was obtained by the NimbleScan software and is derived from the Kolmogorov–Smirnov test comparing the log2 ratios (MIRA versus input) within a 750-bp window centered at each probe and the rest of the data on the array. The asterisks indicate the location of the NEUROD1 target sites.

We randomly selected and validated 11 methylated genomic regions, which were predicted by the array analysis, by using bisulfite-based COBRA assays. The validated targets fell into various major functional categories, including transcription factors and noncoding RNAs such as *GALNTL1, MIR-10A, MIR-129-2, MIR-196A2, MIR-615, MIR-9-3, AMBRA1, HOXD10, PROX1, ZNF672* and *DMRTA2*. Validation of this set of samples revealed the specificity of the analysis. Some of the validated genes are epigenetically altered in various other cancers (*MIR-10A, MIR-129-2, MIR-196A2, HOXD10* and *PROX1*) but other genes have not yet been identified as methylated in any cancer type (*GALNTL1, MIR-615, AMBRA1, ZNF672* and *DMRTA2*). *DMRTA2* methylation was found in 94% of the SCLC tumor patients. The only fact that is known about DMRTA2 is that there is crosstalk of expression with the transcription factor NFIA.[36] Interestingly, there is evidence that NFIA is a key factor for the differentiation of neuronal progenitor cells by downregulating the activity of the Notch signaling pathway via repression of the key Notch effector Hes1.[37] Given the strong enrichment for neuronal differentiation pathways in tumor-specific methylated regions in SCLC (Figure 2) it is tempting to speculate that there is a contribution of DMRTA2 methylation to impaired homeostasis between DMRTA2 and NFIA. There is no functional evidence yet for GALNTL1. These two targets, as well as the many other very frequently methylated genes (Table 2), have the potential to be used as biomarkers for this cancer type.

Gene annotation analysis of tumor-specific promoter methylated targets revealed a substantial subgroup of genes that are specific for neuronal fate commitment, neuronal differentiation and pattern specification processes, along with homeobox and other transcription factors. In comparison, hypermethylated regions in gene bodies showed similar functional enrichment clusters for homeobox genes and pattern specification processes, but significant less enrichment for neuronal fate commitment and for neuronal differentiation, suggesting that the latter functional categories are more specific for promoter-specific methylation. This striking tendency for methylation of neuronal-specific genes may suggest an essential role of this event in SCLC tumor initiation.

Methylation of surrounding proximal promoters is often tightly associated with transcriptional silencing, whereas gene body methylation seems to be associated with transcriptional activation.[27,38] Loss of expression of genes, which are methylated in their proximal promoters, could lead to SCLC tumor initiation. Further studies in this direction will be required to establish experimental evidence. What we do not know at present is whether these genes are unmethylated and expressed in pulmonary neuroendocrine cells and their precursors, the likely cells of origin for SCLC. This specific cell type is currently not available for analysis. This issue does indeed apply to almost all DNA methylation studies done in human cancer to date. The exact