0.66, were set as threshold values for significant differences. A threshold value for the occurrence was set to 70% of all the CRC patient samples in which peptide pairs were detected. In this manner, candidate peptides were selected and further subjected to MS/MS analysis (AXIMA-QIT-TOF; Shimadzu/Kratos) [16]. Proteins were identified by the MASCOT MS/MS Ion Search algorithm (Version 2.0; Matrix Science) using mass lists generated by MASCOT Distiller. The MASCOT search parameters were as follows: trypsin digestion allowing up to two missed cleavages, fixed modifications of 12CNBS (or 13CNBS) and carbamidomethyl (C), variable modifications of oxidation (M), peptide tolerance 0.3 Da and MS/MS tolerance of 0.5 Da. Search results having p-values less than 0.05 were judged as positive identifications.

2.5 Western blot analysis

Total protein extracts (20 μg; CF or CSF) from the tumor and corresponding normal tissue samples of each patient were separated on 10 or 15% SDS-polyacrylamide gels. Proteins were then transferred to a NC membrane and prestained SDS-PAGE standards (Bio-Rad) were used to estimate their molecular weights. The following primary antibodies were used: mouse anti-human Zyxin (ZYX), polyclonal (Abnova, Taipei, Taiwan), mouse anti-human RAN, monoclonal (Abcam, Cambridge, UK), mouse anti-human S-adenosylhomocysteine hydrolase (AHCY), polyclonal (Abnova), rabbit anti-human reticulocalbin 1 (RCN1), monoclonal (Abnova), rabbit anti-human galectin1 (LGALS1), polyclonal (Abcam), and rabbit anti-human Vimentin (VIM), polyclonal (Abcam). NC membranes were incubated with diluted antibody solution for 2 h at room temperature. After washing in PBS, the membranes were incubated at room temperature for 1 h with HRP-conjugated sheep anti-mouse IgG antibody (GE Healthcare) for ZYX, RAN and AHCY, or HRP-conjugated donkey anti-rabbit IgG antibody (GE Healthcare) for RCN1, LGALS1 and VIM. Primary antibody dilutions were anti-ZYX (1:500); anti-RAN (1:1000); anti-AHCY (1:1000); anti-RCN1 (1:1000); anti-LGALS1 (1:1000); and anti-VIM (1: 1000). Secondary antibody dilutions were anti-mouse IgG (1:4000) and anti-rabbit IgG (1:10000). Proteins were then visualized by ECL Plus detection reagents (GE Healthcare), exposed to X-ray film (Kodak, US), and the protein band densities were quantified using "CS Analyzer v3.0" software (ATTO, Tokyo). Used membranes were stained with 0.2% CBB R-250 in 40% MeOH, 10% AcOH for 5 min and destained with 90% MeOH, 2% AcOH for 15 min to confirm equal protein loading and blotting (data not shown).

2.6 Immunohistochemical staining

Ten percent buffered formalin-fixed paraffin-embedded sections were prepared from ten surgically resected cancers. Tissue specimens from the same cancers were also used for proteomics analyses. The streptavidin-biotin immunoperoxidase complex method was used for immunohistochemical analysis. Briefly, 4-μm slices of tissue section were deparaffinized and incubated with 0.03 mol/L citrate buffer (pH 6.0) and heated to 98°C for 40 min for antigen retrieval. Endogenous peroxidase activity was blocked with 0.3% hydrogen peroxide and 0.1% sodium azide in distilled water for 15 min. After three rinses in PBS pH 7.2, 10% bovine serum (Wako, Osaka, Japan) was applied for 10 min to block nonspecific reactions. Sections were incubated with the primary antibody for 60 min at room temperature. Primary antibodies for immunohistochemical staining were the same as those used in the Western blot (WB) analyses. After washing in PBS, the sections were treated with biotinylated sheep anti-mouse IgG (Amersham, London, UK) for ZYX, RAN and AHCY or biotinylated anti-rabbit IgG (Nichirei, Tokyo, Japan) for RCN1, LGALS1 and VIM for 15 min. After washing in PBS, the sections were reacted with streptavidin-biotin peroxidase complex (Dako, Copenhagen, Denmark) at 1:300 dilution for 15 min. The peroxidase reaction was visualized by incubating the sections with 0.02% 3,3'-diaminobenzidine tetrahydrochloride in 0.05 M Tris buffer (pH 7.6) with 0.01% hydrogen peroxide for 3 min. Sections were then counterstained with hematoxylin. Negative control sections were tested using normal mouse or rabbit serum instead of the primary antibody. Tissue sections of normal liver (for AHCY), skin (for VIM), lymph node (for LGALS1) and testis (for RAN, RCN1, and ZYX) were prepared as positive controls according to the manufacturers' recommendations or previous publications. All slides were evaluated by a blinded pathologist. For each immunohistochemical analysis, the mean intensity of the tumor cells or stromal cells was evaluated in comparison with the positive controls as follows: weak, 1+; moderate, 2+; strong, 3+.

3 Results

3.1 Proteomic profiling and identification of differentially expressed proteins in CRC tissues

Differential proteome analysis between CRC and normal tissues from each patient was performed using the NBS method (Fig. 1A). This analysis was performed using CF and CSF samples from each of the 12 patients. After a series of experiments, including NBS labeling, peptide fractionation and MS measurement, 2600–3000 paired peaks were observed per analysis. In this method, the relative ratio of expression for each protein is calculated from the relative ratio of peak intensity (or area) in each pair-peak (NBS-labeled peptides) [12]. Following this relative quantification, pair-peak lists were evaluated (see Section 2) and 320 pairs were judged to have significant differences in protein expression and to occur with significant frequency in patients. After these peaks were subjected to MS/MS analysis, 226 decent MS/MS spectra were obtained, and 156 search results (138 identical peptides) were judged as

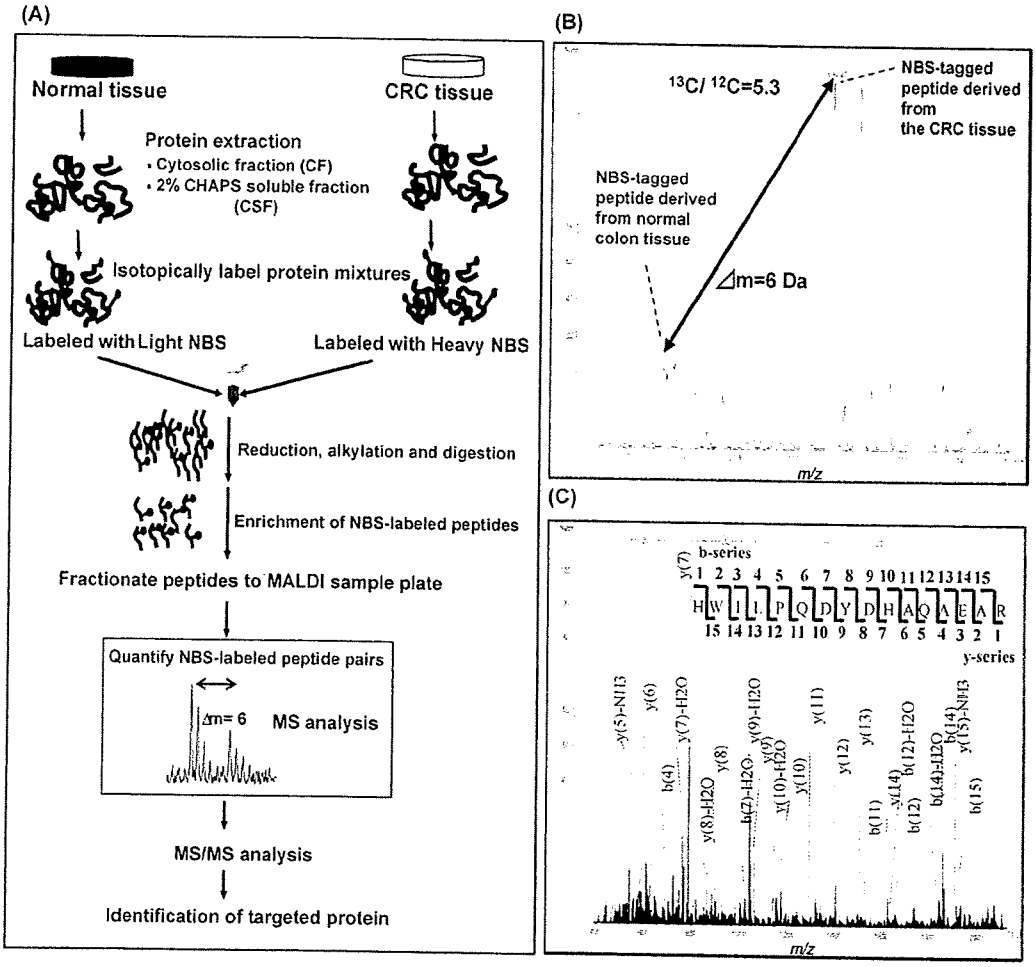


Figure 1. Proteomic analysis of CRC tissue samples by the NBS method. A schematic drawing of NBS analysis (A) and a typical example of MS (B) and MS/MS (C) spectra of an NBS-labeled peptide (HWILPQDYDHAQAEAR) from RCN1 are shown.

positive identifications. This corresponded to 128 proteins, with 71 up-regulated and 57 down-regulated, as listed in Table 1 and Supporting Information 1. Their subcellular localizations are shown in Fig. 2 and Supporting Information 2. The proportion of cytoplasmic proteins identified using CF analysis was nearly half (45.2%), whereas CSF analysis indicated a proportion of less than a quarter (22.2%). In contrast, the proportion of extracellular and plasma membrane proteins identified using CSF analysis (27.0%) was much larger than that identified using CF analysis (16.4%). Most of the proteins were identified from either CF analysis or CSF analysis, and only a few from both (65, 55 and 8 proteins were identified from CF only, CSF only or both fractions, respectively). This means that CF/CSF fractionations were successfully achieved and that these fractionations improved the proteome coverage.

NBS labeling followed by MS analyses was carried out twice for all samples to evaluate experimental variations (Supporting Information 3). Correlation coefficients calculated for each of the 12 patient samples were all above 0.95, indicating that this quantitation method was reliable.

In this study, we focused on 23 up-regulated proteins whose mRNA expression was also up-regulated, as determined by cDNA microarray analysis (unpublished data). We selected six of these proteins (ZYX, RAN, RCN1, AHCY, LGALS1, and VIM) for further characterization, using the criteria that they had not been well studied in relation to CRC and that they have distinct features.

3.2 Verification of results obtained from global quantitative proteome analysis by the NBS method

Although the good reproducibility and reliability of the NBS method have been demonstrated previously, [12, 13, 16] and were demonstrated in the above analyses, we carried out WB analysis to verify our results using an independent method.

Table 1. Differentially expressed proteins in CRC tissues

Protein name	Average T/N ratio [a]	Average Log_2 (T/N ratio) ± SD	Occurrence in patients (%)	Fraction[b]	Previously reported in CRC
Up-regulated proteins in CRC tissues					
Alpha1 acid glycoprotein	2.5	1.3 ± 1.0	9 /11 (81%)	CF	
Alpha 1 acid glycoprotein type 2	NC[c]	NC[c]	10	CF, CSF	
Alpha-tubulin	2.0	1.0 ± 0.54	9 /9 (100%)	CSF	
Beta-tubulin	2.1	1.1 ± 0.36	8 /9 (89%)	CSF	[8]
Apurinic endonuclease	2.0	1.0 ± 0.47	12 /12 (100%)	CF	
Calumenin	2.7	1.4 ± 0.86	10 /10 (100%)	CSF	
Chaperonin1	2.3	1.2 ± 0.46	11 /11(100%)	CSF	[6]
Clathrin heavy polypeptide	1.7	0.78 ± 0.78	7 /9 (89%)	CSF	
Clathrin light polypeptide A	3.2	1.7 ± 1	10/11 (91%)	CSF	
Complement factor H	2.1	1.1 ± 0.33	9 /10 (90%)	CF,CSF	
Cysteine rich intestinal protein 1	1.7	0.78 ± 0.51	8 /9 (89%)	CSF	
Cytokeratin 18	2.8	1.5 ± 0.99	10 /12 (83%)	CSF	[6, 9]
Enolase 1	2.1	1.1 ± 0.92	9 /10 (90%)	CF,CSF	[6, 7]
Ezrin	2.3	1.2 ± 0.6	8 /10 (80%)	CF	[7]
F-box protein 40	2.5	1.3 ± 0.61	11 /12 (91%)	CSF	
Fibrinogen gamma	2.2	1.2 ± 0.5	9 /10 (90%)	CF	
Fk506 Binding Protein 1A	2.2	1.2 ± 0.5	9 /10 (90%)	CF	
Galectin 1	2.1	1.1 ± 0.39	9 /11(81%)	CSF	
Glutathione peroxidase 1	1.8	0.81 ± 0.37	8 /10 (80%)	CF	
Glycyl tRNA synthetase	NC[c]	NC[c]	8	CF	[8]
Glyceraldehyde-3-phosphate dehydrogenase	2.0	1.0 ± 0.82	8 /10 (80%)	CF	
Golgi complex-associated protein 1	2.2	1.1 ± 0.73	8 /9 (89%)	CF	
Heat shock 70kD protein 9B	2.8	1.5 ± 0.85	8 /9 (89%)	CF	
Heat shock protein 27	2.1	1.1 ± 0.34	10 /12 (83%)	CF	
Heparan sulfate proteoglycan 2	2.5	1.3 ± 0.59	8 /9 (89%)	CSF	
Heterogeneous nuclear ribonucleoprotein H2	NC[c]	NC[c]	9	CSF	
High density lipoprotein binding protein	2.1	1.1 ± 0.25	8 /8 (100%)	CSF	
HLA-C	2.1	1.1 ± 0.55	7 /8 (88%)	CF	
Hypothetical protein FLJ38663	2.2	1.1 ± 0.87	7 /9 (78%)	CF	
Inorganic pyrophosphatase	2.5	1.3 ± 0.76	10 /11 (90%)	CF	[6, 8]
Membrane-bound C2 domain-containing protein	2.0	0.98 ± 0.31	8 /9 (89%)	CSF	
Mitogen inducible gene 2 protein	1.9	0.94 ± 0.7	7 /9 (78%)	CF	
6-Phosphogluconolactonase	2.0	0.99 ± 0.78	9 /10 (90%)	CF	
Plastin 2	2.4	1.2 ± 0.49	10 /10 (100%)	CF	
Plectin 1	2.0	1.0 ± 0.8	8 /10 (80%)	CSF	
Proteasome subunit p58	2.1	1.1 ± 0.4	8 /8 (100%)	CSF	
Protein tyrosine phosphatase, receptor type c	NC[c]	NC[c]	8	CSF	
Protein tyrosine phosphatase, receptor type, α	NC[c]	NC[c]	8	CSF	
Pyruvate kinase 3	1.9	0.93 ± 0.4	11 /12 (92%)	CF	[6]
RAB18, member RAS oncogene family	2.7	1.5 ± 0.86	10 /12 (83%)	CSF	
RAB22A	1.9	0.94 ± 0.73	7 /10 (70%)	CSF	
RACK1	2.0	1.0 ± 0.32	10 /10 (100 %)	CF	
Radixin	1.8	0.84 ± 0.76	8 /12 (73%)	CF	
RAN, member RAS oncogene family	2.0	0.99 ± 0.67	9 /11 (81%)	CF	
Reticulocalbin 1	3.4	1.8 ± 0.96	9 /10 (90%)	CF	
Rhodanese; thiosulfate sulfurtransferase	1.9	0.95 ± 0.66	7 /10 (70%)	CF	
Ribonuclease RNase A family 3	NC[c]	NC[c]	8	CSF	
Ribosomal protein L13	3.4	1.8 ± 0.98	10 /10 (100%)	CSF	
Ribosomal protein L27a	2.1	1.0 ± 0.71	8 /11 (73%)	CSF	
Ribosomal protein L4	2.0	1.0 ± 0.64	9 /11 (82%)	CSF	
Ribosomal protein S18	2.8	1.5 ± 0.46	10 /10 (100%)	CSF	
Ribosomal protein S29	2.0	1.0 ± 0.47	7 /8 (88%)	CSF	
Ribosome binding protein 1	1.8	0.87 ± 0.34	10 /11(91%)	CF	[7]
S adenosylhomocysteine hydrolase	2.3	1.2 ± 0.79	10 /11 (90%)	CF	

Table 1. Continued

Protein name	Average T/N ratio [a]	Average Log_2 (T/N ratio) ± SD	Occurrence in patients (%)	Fraction[b]	Previously reported in CRC
S100 calcium binding protein A9	2.2	1.2 ± 0.7	9/ 11 (82%)	CF	[9]
Solute carrier family 25, member 5	2.2	1.1 ± 0.7	8 /9 (89%)	CSF	
Solute carrier family 3, member 2	2.0	1.0 ± 0.31	8 /9 (89%)	CSF	
Splicing factor 3B, subunit 3	2.8	1.5 ± 0.99	9 /10 (90%)	CF	[7]
Splicing factor, arginine/serine-rich 3 (SRp20)	2.3	1.2 ± 0.52	7 /8 (88%)	CF	
TLS protein	NC[c]	NC[c]	9	CSF	
Transgelin	2.1	1.1 ± 0.78	7 /9 (78%)	CF	[8, 10]
Transgelin 2	2.3	1.2 ± 0.32	10 /10 (100%)	CF	[6]
Triosephosphate isomerase 1	2.0	0.99 ± 0.92	10 /12 (83%)	CF, CSF	[6, 8]
Tumor rejection antigen 1	NC[c]	NC[c]	8	CSF	[6]
Ubiquitin activating enzyme 1	2.1	1.1 ± 0.47	9 /10 (90%)	CF	
Ubiquitin isopeptidase T	2.3	1.2 ± 0.6	9 /10 (90%)	CSF	
U5 snRNP-specific protein, 116 kDa	2.5	1.3 ± 0.71	9 /11 (82%)	CSF	
Vimentin	2.5	1.3 ± 0.39	9 /9 (100%)	CSF	[6]
Vitronectin	2.1	1.1 ± 0.8	7 /8 (88%)	CSF	
XTP3 transactivated protein A	2.4	1.3 ± 0.52	9 /10 (90%)	CF	
Zyxin	2.2	1.1 ± 0.7	9 /10 (90%)	CF	

a) An average (T/N ratio) is calculated by exponential transformation of an average Log_2 (T/N ratio).
b) Fraction; CF: cytosolic fraction, CSF: 2% CHAPS-soluble fraction.
c) NC; not calculated. This is because a protein is exclusively detected in CRC (or normal) tissues.

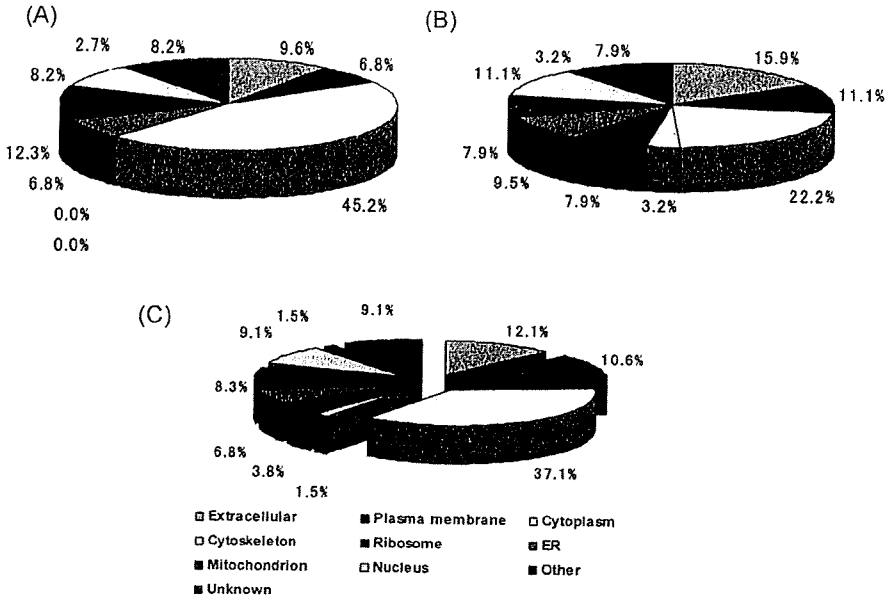


Figure 2. Classification of differentially expressed proteins in CRC tissues. Identified proteins that were differentially expressed in CRC tissues are classified into categories based on their subcellular localization. Proteins identified in the cytosolic fraction (CF) (A), and those identified in the 2% CHAPS-soluble fraction (CSF) (B) and the total of 128 proteins (C) are represented graphically.

Using the same fraction (CF or CSF) and the same CRC patient samples used in the NBS analyses, WB analyses were performed for the six selected proteins (Fig. 3). Generally, high T/N ratios were again observed in WB analyses, though the precise ratio was sometimes unmatched in one-to-one data comparison with the NBS analysis. The inconsistency sometimes observed between NBS and WB data may be due to the different assay methods (see Section 4). Because the high T/N ratios observed in the NBS analyses were confirmed by WB analyses for all six selected proteins, these proteins were further evaluated by immunohistochemical staining analysis.

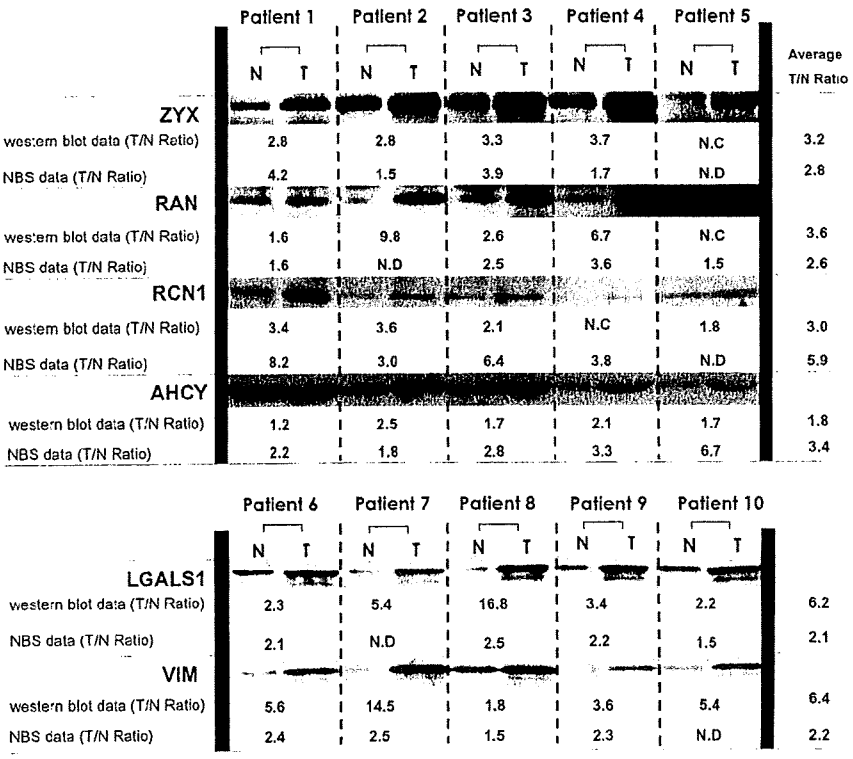


Figure 3. Western blot (WB) analysis for data verification. WB analyses were performed using matched pairs of normal (N) and tumor (T) tissues to detect the following proteins: ZYX, RAN, RCN1, AHCY, LGALS1, and VIM. CF samples were used to detect the four proteins in the upper panel and CSF samples were used to detect the two proteins in the lower panel. The T/N ratios obtained from WB and NBS analyses are shown below each blot. The average T/N Ratio for each protein was calculated from the ratio in each sample where both NBS and WB data were obtained. N.C.: not calculated. N.D.: not detected.

3.3 Further validation by immunohistochemical staining

Finally, immunohistochemical staining was performed to investigate the localization of the six selected proteins and to examine their expression patterns in whole CRC tissues (Fig. 4 and Supporting Information 4). All of these proteins, in all samples tested, were immunohistochemically detected more frequently in CRC tissues than in adjacent normal tissues. Cancer cells expressed five proteins (ZYX, RAN, RCN1, AHCY, and LGALS1); RCN1 and AHCY were strongly detected and ZYX, RAN and LGALS1 were found at moderate levels, on average, in all samples tested, although their expression patterns were not homogenous. In particular, RAN, RCN1 and LGALS1 were localized partially in cancer cells. On the other hand, normal colorectal epithelial cells expressed only ZYX, RCN1 and AHCY, and their expression was generally weak. Various stromal cells were positive for five proteins (ZYX, RAN, AHCY, LGALS1, and VIM), including leukocytes, blood vessels, nerves and fibroblasts. VIM was strongly and specifically expressed in stromal cells close to the cancer cells. ZYX and LGALS1 were moderately expressed in stromal cells generally, and their expression here was as strong as that of cancer cells. Stromal cells close to the cancer cells expressed these three proteins more intensely than those distant from cancer cells. Five of the six selected proteins (ZYX, RAN, AHCY, LGALS1, and VIM) were expressed not only in cancer cells, but also in the surrounding stromal cells.

4 Discussion

We performed proteomic profiling of CRC tissue samples to identify novel biomarker candidates by the NBS method. Proteins that were differentially expressed between tumor and normal tissues, with significant differences in expression and affecting most of the patients, were selected and a final set of 128 proteins was identified. Of these, 23% (30 proteins) were reported in earlier proteomic studies on CRC [6–10]. Interestingly, the present study has led to identification of about 100 novel CRC-associated proteins that have not been reported to associate with CRC before.

In this study, two fractions (CF and CSF) were prepared in order to increase the range of analyses. Many extracellular and plasma membrane proteins were recovered simultaneously and identified in CSF fractions. CEA, a well-known biomarker used for clinical detection of CRC, is a GPI-anchored membrane glycoprotein. It has been suggested that CEA is cleaved by glycosylphosphatidylinositol-phospholipase D (GPI-PLD) and then secreted into blood [19]. Thus, plasma-membrane proteins have the best chance of being secreted into circulatory systems, along with extracellular proteins. In other words, plasma-membrane proteins and extracellular proteins are excellent candidates for biomarker development. Viewed in this context, our report seems to contain many potential CRC biomarker candidates.

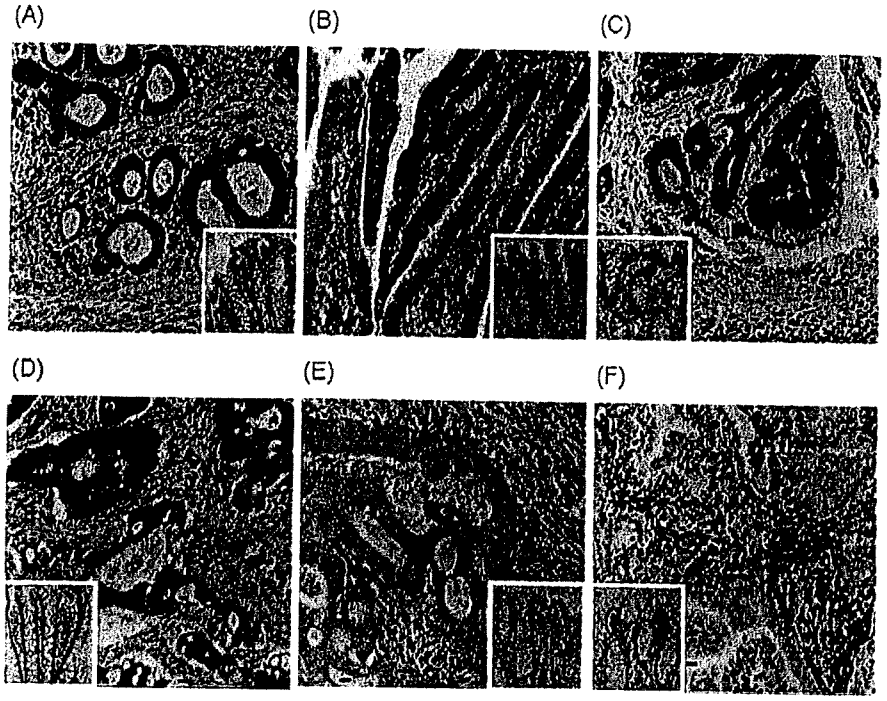


Figure 4. Immunohistochemical staining (IHC) for the six selected proteins in CRC tissues. Insets show IHC staining of normal colonic tissue. Bars indicate 20 μm in all figures, including insets (original magnification ×140, insets ×70). Each displayed picture (A–F) is a typical example of ten tested samples. (A) ZYX was expressed in cancer cells as well as in stromal cells close to cancer cells, including leukocytes and blood vessels. Weak ZYX expression was observed in normal epithelial cells and in stromal cells distant from cancer cells (Inset). (B) RAN was expressed in cancer cells, while normal epithelial cells were negative. Some leukocytes adjacent to cancer cells also expressed RAN. (C) RCN1 was expressed in cancer cells and weakly in normal epithelial cells. No RCN1 expression was detected in stromal cells. (D) AHCY was expressed in cancer cells and weakly in normal epithelial cells. Weak AHCY expression was observed in stromal cells such as leukocytes and blood vessels. (E) LGALS1 was expressed in stromal cells including leukocytes, nerve cells and fibroblasts. Expression in stromal cells adjacent to cancer cells was much stronger. Some cancer cells expressed LGALS1, while normal epithelial cells were negative. (F) VIM was strongly expressed in stromal cells including leukocytes, fibroblasts and blood vessels. Expression in stromal cells adjacent to cancer cells was much stronger. Neither normal epithelial cells nor cancer cells expressed VIM at all.

There are some discrepancies between the NBS and WB results (Fig. 3; for example, AHCY for Patient 5, LGALS1 for Patient 8, VIM for Patient 7). This probably was due to differences in the two analytical methods. In WB analysis, whole proteins, including PTM or alternative splicing variants, can be detected by a shift in migration, although subtle differences of molecular weight change caused by a small modification or a point mutation are overlooked. On the other hand, in NBS analysis, peptides are detected with a resolution power of less than 1 Da, although PTM, mutations, and splice variants not present in the NBS-labeled peptides will be overlooked. Thus, the NBS method by itself is not suitable for a comprehensive analysis of PTM moieties of proteins because the coverage of each protein is quite low. This is indeed a drawback and a limitation of the NBS method, which must be compensated by other methods. However, global proteome analyses with respect to PTM-proteins can be performed by the NBS method if it is combined with prior enrichment of the sample [18].

Six selected proteins showed high expression levels in CRC tissues and occurred with high frequency in CRC patients [ZYX 90% (9/10), RAN 81% (9/11), RCN1 90% (9/10), AHCY 90% (10/11), and LGALS1 81% (9/11), respectively]. These proteins were also selected because they have not been well studied in relation to CRC, and so they were considered to be novel biomarker candidates. Each of these is discussed in detail below.

ZYX is a zinc-binding phosphoprotein that is a member of the LIM protein family. This protein is widely expressed in human tissues and is most prominent in lung and colon tissues [20]. It has been suggested that ZYX might enhance hepatocellular carcinoma cell migration and intravasation through its action on the actin cytoskeleton, as it promotes cell dissemination as a part of integrin-signaling pathways [21]. However, there have been no reports demonstrating an association of ZYX with CRC. In this study, we have demonstrated high expression of ZYX in CRC tissues for the first time. Moreover, ZYX was highly expressed in both cancer cells and in the surrounding stromal cells. Because this

molecule is also strongly expressed in blood vessels in close proximity to cancer cells, it might be secreted into the blood of CRC patients.

RAN is a small GTP-binding protein belonging to the RAS superfamily. This molecule has also proven essential for various mammalian cellular processes, such as nuclear-cytoplasmic transport, cell cycle progression, nuclear organization, nuclear envelope assembly, mitotic control and genomic instability [22, 23]. Because genetic instability is a major factor in carcinogenesis and development of cancer, this protein is considered associated with carcinogenesis. It has been reported that high-level expression of RAN is strongly associated with the prognosis of epithelial ovarian tumors [24]. In this study, we demonstrated its up-regulation in both cancer cells and in the surrounding stromal cells. Taking into consideration these data and previous reports, it appears that this protein might be a key molecule in CRC carcinogenesis.

RCN1 is a calcium-binding protein expressed in the ER. This protein contains six repeats of a domain containing an EF-hand motif, which is considered to play a role in Ca^{2+}-dependent cell adhesion [25] This protein regulates cadherin expression in breast carcinoma and colorectal carcinoma cells (SW480) [26, 27], and high expression of this protein in hepatoma cells has been demonstrated [28]. We noted for the first time high expression levels of this protein in CRC tissue. In addition, we demonstrated that RCN1 was overexpressed with high frequency only in cancer cells. Our data suggest that RCN1 might be a promising candidate for biomarker development.

AHCY catalyzes hydrolysis of S-adenosylhomocysteine to adenosine and homocysteine. This enzyme is crucial in the control of transmethylation reactions, and a deficiency of this molecule induces hypermethioninemia [29]. There is little evidence indicating any association of this molecule with cancer. Another proteomics study demonstrated up-regulation of AHCY in CRC tissue [8]; however, this was not characterized further. The present study utilized IHC to show, for the first time, high expression of AHCY in cancer cells and weaker expression in the surrounding stromal cells.

Galectins regulate pleiotropic biological functions involved in cell growth, differentiation, adhesion, RNA processing, apoptosis and malignant transformation [30]. Galectin-1 is found extracellularly in many tissues in both normal and pathological conditions, and several reports have demonstrated an association of galectin-1 with cancer [30]. Up-regulation of LGALS1 in CRC tumor tissue was found using another proteomic approach [31]; however, no follow-up studies have been reported to date. In the present study, this protein was strongly expressed in stromal cells adjacent to cancer cells, in addition to its up-regulation in tumor tissue.

VIM is an intermediate filament that represents a third class of well-characterized cytoskeletal elements, along with the actins and tubulins. This protein is considered a key protein in cell physiology, cellular interactions, and organ homeostasis [32]. In previous reports, IHC examination has shown that most cases of primary colonic malignant melanoma, which is a rare tumor, were positive for VIM [33]. In the present study, this molecule was strongly expressed in stromal cells in the vicinity of cancer cells, and was associated with CRC.

Analyses of whole tissues including stroma are important to understand cancer biology and to discover novel biomarker candidates. This is because as much as a half of tumor tissues are composed of stroma, and cancer cells are frequently influenced by this. In addition, stroma cells may be influenced by neighboring cancer cells, and changes of the stroma could constitute a "cancer signal". IHC experiments in this study revealed that two proteins (LGALS1 and VIM) were up-regulated in CRC tissues, but they are located primarily in stroma, not in tumor cells. This observation supports the importance of whole tissue analysis. In addition, it suggests that a proteomic study should be compensated and validated by IHC, because it alone does not reveal detailed information regarding protein localization, which is required for precise functional analyses.

In this study, proteins up-regulated in CRC tissues, whose mRNA expression was also up-regulated, were primarily selected for further studies. It is natural to focus on up-regulated proteins in CRC to identify potential diagnostic biomarkers to achieve early detection of CRC. This set of proteins may include a key protein that is responsible for carcinogenesis. Considering the above possibility, genes with high mRNA expression in CRC are the best targets for gene therapy, such as RNA interference. This is why we prioritized investigating these proteins over others in this study. However, up-regulated proteins that showed a negative correlation with mRNA expression require further investigation because they were detected only by proteomic studies.

To identify biomarker candidates for early diagnosis, examinations of specimens from patients diagnosed at an early stage of disease may be ideal. However, advanced tumor specimens were used in this study for the first screening, as previously reported [6–10], partly because it is difficult to obtain enough early stage specimens. Although many proteins identified in this experiment may not be applicable to early diagnosis, several of them are anticipated to be already altered and to be applicable as early stage markers, as proved in earlier studies [7, 9]. This is why proteins identified from advanced stage specimens may still be used as early diagnostic markers. We are now examining whether selected candidate proteins are detected at significantly higher levels in sera of CRC patients in comparison with healthy volunteers.

We originally assumed that the term "biomarker" is simply used in clinical diagnostic scenes, but it could be used in clinicopathological ones as well. For example, there are few cases in which the pathological diagnosis for cancer is difficult. However, it is sometimes difficult to specify

the localization of the primary tumor when a metastatic focus is first detected in cancer screening [34]. In such a case, the proteins identified in CRC tissues in this study would be very useful to identify the primary site, if they were applicable.

We believe that the present study will contribute to future improvements in diagnostic/prognostic applications, understanding of CRC carcinogenesis, and the discovery of new therapeutic targets and drugs.

We thank Dr. Koichi Tanaka, Dr. Susumu Iwasa, and Dr. Susumu Tsunasawa for excellent technical advice and useful discussions.

The authors have declared no conflict of interest.

5 References

[1] Carpelan-Holmstrom, M. A., Haglund, C. H., Roberts, P. J., Differences in serum tumor markers between colon and rectal cancer. Comparison of CA 242 and carcinoembryonic antigen. *Dis. Colon Rectum.* 1996, *39*, 799–805.

[2] Von Kleist, S., Hesse, Y., Kananeeh, H., Comparative evaluation of four tumor markers, CA 242, CA 19/9, TPA and CEA in carcinomas of the colon. *Anticancer Res.* 1996, *16*, 2325–2332.

[3] Carpelan-Holmstrom, M. A., Haglund, C. H., Javinen, H. J., Roberts, P. J. Serum CA 242 and CEA detect different patients with recurrent colorectal cancer. *Anticancer Res.* 1996, *16*, 981–986.

[4] Takemasa, I., Higuchi, H., Yamamoto, H., Sekimoto, M. *et al.*, Construction of preferential cDNA microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. *Biochem. Biophys. Res. Commun.* 2001, *285*, 1244–1249.

[5] Komori, T., Takemasa, I., Higuchi, H., Yamasaki, M. *et al.*, Identification of differentially expressed genes involved in colorectal carcinogenesis using a cDNA microarray. *J. Exp. Clin. Cancer Res.* 2004, *23*, 521–527.

[6] Tomonaga, T., Matsushita, K., Yamaguchi, S., Oh-Ishi, M. *et al.*, Identification of altered protein expression and post-translational modifications in primary colorectal cancer by using agarose two-dimensional gel electrophoresis. *Clin. Cancer Res.* 2004, *10*, 2007–2014.

[7] Roessler, M., Rollinger, W., Palme, S., Hagmann, M. *et al.*, Identification of nicotinamide N-methyltransferase as a novel serum tumor marker for colorectal cancer. *Clin. Cancer Res.* 2005, *11*, 6550–6557.

[8] Friedman, D. B., Hill, S., Keller, J. W., Merchant, N. B. *et al.*, Proteome analysis of human colon cancer by two-dimensional difference gel electrophoresis and mass spectrometry. *Proteomics* 2004, *4*, 793–811.

[9] Stulik, J., Hernychova, L., Porkertova, S., Knizek, J. *et al.*, Proteome study of colorectal carcinogenesis. *Electrophoresis* 2001, *22*, 3019–3025.

[10] Bi, X., Lin, Q., Foo, T. W., Joshi, S. *et al.*, Proteomic Analysis of Colorectal Cancer Reveals Alterations in Metabolic Pathways. *Mol. Cell. Proteomics* 2006, *5*, 1119–1130.

[11] Rabilloud, T., Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains, *Proteomics* 2002, *2*, 3–10.

[12] Kuyama, H., Watanabe, M., Toda, C., Ando, E. *et al.*, An approach to quantitative proteome analysis by labeling tryptophan residues. *Rapid Commun. Mass Spectrom.* 2003, *17*, 1642–1650.

[13] Matsuo, E., Toda, C., Watanabe, M., Iida, T. *et al.*, Improved 2-nitrobenzenesulfenyl method: optimization of the protocol and improved enrichment for labeled peptides. *Rapid Commun. Mass Spectrom.* 2006, *20*, 31–38.

[14] Matsuo, E., Toda, C., Watanabe, M., Ojima, N. *et al.*, Selective detection of 2-nitrobenzenesulfenyl-labeled peptides by matrix-assisted laser desorption/ionization-time of flight mass spectrometry using a novel matrix. *Proteomics* 2006, *6*, 2042–2049.

[15] Cagney, G., Amiri, S., Premawaradena, T., Lindo, M. *et al.*, *In silico* proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci.* 2003, *13*, 5.

[16] Iida, T., Kuyama, H., Watanabe, M., Toda, C. *et al.*, Rapid and efficient MALDI-TOF MS peak detection of 2-nitrobenzenesulfenyl labeled peptides using the combination of HPLC and an automatic spotting apparatus. *J. Biomol. Tech.* 2006, *17*, 333–341.

[17] Ou, K., Kesuma, D., Ganesan, K., Yu, K. *et al.*, Quantitative profiling of drug-associated proteomic alterations by combined 2-nitrobenzenesulfenyl chloride (NBS) isotope labeling and 2DE/MS identification. *J. Proteome Res.* 2006, *5*, 2194–2206.

[18] Ueda, K., Katagiri, T., Shimada, T., Irie, S. *et al.*, Comparative profiling of serum glycoproteome by sequential purification of glycoproteins and 2-nitrobenzensulfenyl (NBS) stable isotope labeling: A new approach for the novel biomarker discovery for cancer. *J. Proteome Res.* 2007, *6*, 3475–3483.

[19] Naghibalhossaini, F., Ebadi, P., Evidence for CEA release from human colon cancer cells by an endogenous GPI-PLD enzyme. *Cancer Lett.* 2006, *234*, 158–167.

[20] Macalma, T., Otte, J., Hensler, M. E., Bockholt, S. M. *et al.*, Molecular characterization of human zyxin. *J. Biol. Chem.* 1996, *271*, 31470–31478.

[21] Sy, S. M., Lai, P. B., Pang, E., Wong, N. L. *et al.*, Novel identification of zyxin upregulations in the motile phenotype of hepatocellular carcinoma. *Modern Pathol.* 2006, *19*, 1108–1116.

[22] Ren, M., Drivas, G., D'Eustachio, P., Rush, M. G. Ran/TC4: a small nuclear GTP-binding protein that regulates DNA synthesis. *J. Cell Biol.* 1993, *120*, 313–323.

[23] Di Fiore, B., Ciciarello, M., Lavia, P. Mitotic functions of the Ran GTPase network: the importance of being in the right place at the right time. *Cell Cycle* 2004, *3*, 305–313.

[24] Ouellet, V., Guyot, M. C., Le Page, C., Filali-Mouhim, A. *et al.*, Tissue array analysis of expression microarray candidates identifies markers associated with tumor grade and outcome in serous epithelial ovarian cancer. *Int. J. Cancer* 2006, *119*, 599–607.

[25] Tachikui, H., Navet, A. F., Ozawa, M. Identification of the $Ca^{(2+)}$-binding domains in reticulocalbin, an endoplasmic reticulum resident $Ca^{(2+)}$-binding protein with multiple EF-hand motifs. *J. Biochem.* 1997, *121*, 145–149.

[26] Liu, Z., Brattain, M. G., Appert, H. Differential display of Reticulocalbin in the highly invasive cell line, MDA-MB-435, versus the poorly invasive cell line. MCF-7. *Biochem. Biophys. Res. Commun.* 1997, *231*, 283–289.

[27] Vermeulen, S. J., Bruyneel, E. A., Bracke, M. E., De Bruyne, G. K. *et al.*, Transition from the non-invasive to the invasive phenotype and loss of α-catenin in human colon cancer cells. *Cancer Res.* 1995, *55*, 4122–4728.

[28] Yu, L.R., Zeng, R., Shao, X.X., Wang, N. *et al.*, Identification of differentially expressed proteins between human hepatoma and normal liver cell lines by two-dimentional electrophoresis and liquid chromatography-ion trap mass spectrometry. *Electrophoresis* 2000, *21*, 3058–3068.

[29] de la Haba, G., Cantoni, G. L. The enzymatic synthesis of *S*-adenosyl-L-homocysteine from adenosine and homocysteine. *J. Biol. Chem.* 1959, *234*, 603–608.

[30] Perillo, N. L., Marcus, M. E., Baum, L. G., Galectins:versatile modulators of cell adhesion, cell proliferation, and cell death. *J. Mol. Med.* 1998, *76*, 402–412.

[31] Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A. *et al.*, A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1920–1932.

[32] Ivaska, J., Pallari, H. M., Nevo, F. J., Eriksson, J. E., Novel functions of vimentin in cell adhesion, migration, and signaling. *Exp. Cell Res.* 2007, *313*, 2050–2062.

[33] Mori, D., Satoh, T., Nakafusa, F., Tanaka, M. *et al.*, Primary colonic malignant melanoma. *Pathol. Int.* 2006, *56*, 744–748.

[34] Pavilidis, N., Cancer of unknown primary: biological and clinical characteristics. *Ann. Oncology*, 2003, *14*, 11–18.

RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2008; **22**: 1109–1116
Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/rcm.3482

RCM

Sensitive detection of phosphopeptides by matrix-assisted laser desorption/ionization mass spectrometry: use of alkylphosphonic acids as matrix additives

Hiroki Kuyama[1], Kazuhiro Sonomura[2] and Osamu Nishimura[1,2]*

[1]Institute for Protein Research, Osaka University, Suita 565-0871, Japan
[2]Life Science Laboratory, Shimadzu Corporation, Kyoto 604-8511, Japan

Received 14 December 2007; Revised 31 January 2008; Accepted 3 February 2008

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) has been one of the most powerful tools for analyzing protein phosphorylation. However, it is frequently difficult to detect phosphopeptides with high sensitivity by MALDI-MS. In our investigation of matrix/matrix-additive substances for improving the phosphopeptide ion response in MALDI-MS, we found that the addition of low-concentration alkylphosphonic acid to the matrix/analyte solution significantly enhanced the signal of phosphopeptides. In this study, the combination of methanediphosphonic acid and 2,5-dihydroxybenzoic acid gave the best results. In addition to enhancing the signal of the phosphopeptides, alkylphosphonic acid almost completely eliminated the signals of sodium and potassium ion adducts. We report herein sensitive detection of phosphopeptides by MALDI-MS with the use of alkylphosphonic acids as matrix additives. Copyright © 2008 John Wiley & Sons, Ltd.

Phosphorylation is one of post-translational modifications in proteins, which is a ubiquitously found biological event in living systems.[1–7] One-third of all eukaryotic proteins are reported to be phosphorylated,[8,9] and phosphorylation plays a pivotal role in a wide range of important signal transduction pathways and other cellular processes such as growth, metabolism, proliferation, motility, interaction, and differentiation in a living cell.[1–7] Hence, protein phosphorylation has been recognized as one of the most relevant post-translational modifications, and localization of phosphorylation sites in the protein sequence is an important goal for understanding regulation mechanisms.

Although the importance has been highly acknowledged, the characterization of the protein phosphorylation sites in biologically derived proteins is still challenging.

Since its introduction more than 20 years ago, matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) has become one of the most powerful tools for analyzing proteins, peptides, oligosaccharides, and so on, because of its high sensitivity, robustness and easy sample preparation. Hence, phosphorylation has frequently been analyzed by MALDI-MS.

However, although as pointed out above MALDI-MS has many advantages, a serious problem still remains which hampers the analysis of phosphopeptides by MALDI-MS. This is because phosphorylated peptides are frequently more difficult to detect and analyze by MALDI-MS than their non-phosphorylated cognates. The difficulty of detecting phosphorylated peptides arises from their low abundance and intrinsic low ionizability. In addition, ionization is often suppressed in the presence of non-phosphopeptides as in the normal proteolytic digest of a phosphoprotein.

To remove impediments to analysis, there have been, in general, three typical workflows for protein phosphorylation analysis, chemical derivatization,[10–13] selective enrichment,[14,15] and the choice of matrix/matrix-additive system.[16–20] Although these three approaches each have their own advantages, the choice of matrix/matrix additive to enhance the response of phosphopeptides in MALDI-MS is a very appealing approach because no chemical or enzymatic treatments are required nor are selective chromatographic methods necessary.

A few years ago, we reported a versatile dephosphorylation procedure using aqueous HF solution or HF-pyridine to detect phosphopeptides in a digestive mixture,[21] and it has since been used by several groups.[22–25] It is a very efficient and robust measure for cleaving the phosphomonoester linkage in phosphopeptides[22,23,25,26] and phosphoproteins,[24] and facilitates the detection of phosphorylated peptides in a digestive mixture. Comparison of the spectra obtained before and after the treatment with aqueous HF solution or HF-pyridine clearly demonstrates the mass shift (−80 Da) due to removal of the phosphate group. However, phosphorylated peptides themselves are sometimes not discernible in a MALDI mass spectrum because of their low concentration and/or their low ionizability.

**Correspondence to*: O. Nishimura, Institute for Protein Research, Osaka University, Suita 565-0871, Japan.
E-mail: osamu_nishimura@protein.osaka-u.ac.jp

RCM

One of our goals has been to directly detect phosphorylated peptides by MALDI-MS. So far, there is no universal methodology for the sensitive detection of phosphopeptides. The goal is to analyze phosphopeptides with high sensitivity and resolution, and it is better to do so using a less cumbersome, less time-consuming approach. Therefore, we chose the last approach (the choice of matrix/matrix additive system) from the viewpoint of readiness of operation, and we undertook this study for more efficient detection of phosphopeptides by MALDI-MS.

Based on the finding of the enhanced response of phosphopeptides by the addition of strong acid,[27] Jensen *et al.* reported that phosphoric acid (PA) was proved to be the matrix additive of choice for the sensitive detection of phosphopeptides.[17] PA works well in enhancing the phosphopeptide signal. However, with use of PA as an additive, some phosphopeptides as well as non-phosphopeptides escape detection by MALDI-MS in the negative ion mode.[17]

This drove us to investigate a more efficient and reliable matrix additive for detecting phosphopeptides by MALDI-MS. In a previous report on matrix selection,[28] we found that the structural similarity between the analyte and matrix might be the key for the proper choice of matrix and/or matrix additive. With this in mind, we started investigating organic and inorganic phosphorous acid as a matrix additive for the enhanced detection of phosphopeptides.

EXPERIMENTAL

Materials

Ovalbumin(chicken egg white), α-casein (bovine milk), albumin (bovin serum; BSA), phosphorylase b (rabbit muscle) and lysozyme (chicken egg white) were purchased from Sigma (St. Louis, MO, USA). Sequencing-grade modified trypsin was obtained from Promega (Madison, WI, USA). High-purity MALDI-MS grade 2,5-dihydroxybenzoic acid (DHBA) and α-cyano-4-hydroxycinnamic acid (CHCA) were obtained from Shimadzu GLC (Tokyo, Japan). BPNA, MPNA, SAA, PNAA and EDTA were purchased from Sigma (St. Louis, MO, USA). The abbreviations cited here indicate matrix additive candidates (see Fig. 1). PNA was purchased from Wako Pure Chemical Industries, Ltd. (Osaka, Japan). PA and MLA were obtained from NACALAI TESQUE, Inc. (Kyoto, Japan). MDPNA, EDSA, NTA, NTMPNA and EDTMPNA were obtained from Tokyo Chemical Industry Co., Ltd. (Tokyo, Japan). EHDPNA and EDPNA were purchased from Alfa Aesar (UK). SCA was obtained from Fluka Chemie AG (Switzerland). Peptides (1P, 2P, 3P and 3PPP; see Table 1) were purchased from AnaSpec, Inc. (San Jose, CA, USA). WAGGDASGE and WAGGDApSGE were purchased from American Peptide Company, Inc. (Sunnyvale, CA, USA). TSTEPQYQPGENL and TSTEPQpYQPGENL were purchased from BACHEM AG (Switzerland). GFETVPETG-NH_2 and GFETVPEpTG-NH_2 were synthesized in-house using a model PSSM-8 peptide synthesizer (Shimadzu) by the Fmoc strategy. All other chemicals were analytical reagent grade and used without further purification.

Sample preparation

Three matrix-additive candidates (NTA, EDTA, and EDTMPNA) were dissolved with water to yield saturated solutions. PA and NTMPNA were dissolved with water to give 3% and 1% solutions, respectively (v/v). Other matrix-additive candidates were dissolved in water to 1% (w/w). The matrix solution was prepared by dissolving 5 mg of DHBA in 0.5 mL of 50% aqueous acetonitrile. CHCA solution, saturated in the same solution system as in DHBA, was used as a matrix for the external calibration.

Model peptides were dissolved in water to the concentrations used.

Tryptic digestion was performed in 50 mM Tris-HCl (pH 7.8), 5 mM $CaCl_2$ at an enzyme-to-substrate ratio of 1:20 (w/w) at 37°C overnight. The resulting digests were used with an appropriate dilution.

MALDI-MS

An AXIMA-CFR plus mass spectrometer (Shimadzu/Kratos, Manchester, UK) was used to obtain all MALDI-TOF mass spectra. The operating conditions were as follows: nitrogen laser (337 nm); reflectron mode; positive or negative mode. The accelerating voltage in the ion source was 20 kV. A standard stainless steel target plate was used for the analysis. A portion (0.3 μL) of each analyte, matrix, and matrix-additive solution was mixed on the target plate and analyzed after drying. The *m/z* values in the spectra were externally calibrated with angiotensin II (human) and ACTH fragment 18–39 (human) using CHCA as a matrix. All measurements were repeated in at least three independent experiments.

RESULTS AND DISCUSSION

Screening of matrix additives

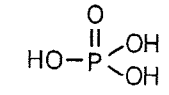
We used 16 different compounds varying from inorganic to organic acids for screening, as summarized in Fig. 1. To evaluate these compounds as matrix additives, a peptide mixture containing four different phosphopeptides in equimolar amount (final concentration of each peptide was 0.1 pmol/μL) was prepared. The contents were three monophosphorylated peptides, and one triphosphorylated peptide, ranging from *m/z* 1438 to 1880, as listed in Table 1.

We used DHBA, which is known to be a 'cool' matrix, as a matrix in this study. The formed ions remain intact during MALDI mass analysis because they have low internal energy when using DHBA as a matrix. Therefore, DHBA is frequently used for phosphopeptide analysis in MALDI-MS.

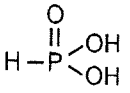
The signal intensities of these peptides are listed in Table 2 which were measured by MALDI-MS using DHBA as a matrix with or without the candidate compound as an additive. In this data acquisition, the laser fluence was set identical except for PA and PNA. When using PA or PNA as an additive, higher laser fluence was needed to produce molecular ions.

PA has an enhancing effect for phosphopeptides, and its cognate phosphonic acid (PNA) is comparably effective. Among acidic groups of COOH, SO_3H, and $P(=O)(OH)_2$, the phosphonic acid group is apparently responsible for the enhancement as determined from the comparisons in the group of (MLA, SAA, PAA) and (SCA, EDSA, EDPNA). This

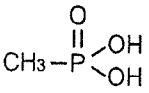
DOI: 10.1002/rcm



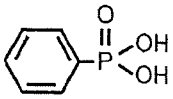
Phosphoric acid (PA)



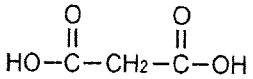
Phosphonic acid (PNA)



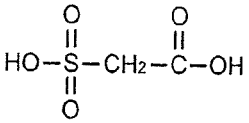
Methanephosphonic acid (MPNA)



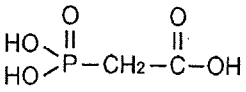
Benzenephosphonic acid (BPNA)



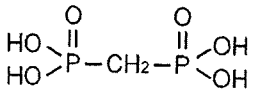
Malonic acid (MLA)



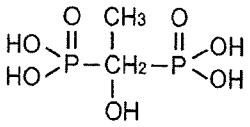
Sulfoacetic acid (SAA)



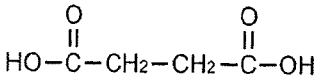
Phosphonoacetic acid (PNAA)



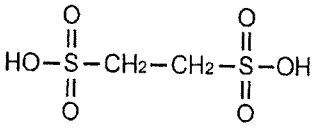
Methanediphosphonic acid (MDPNA)



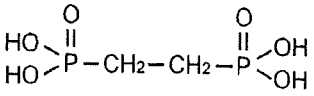
Ethane-1-hydroxy-1,1-diphosphonic acid (EHDPNA)



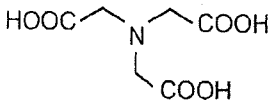
Succinic acid (SCA)



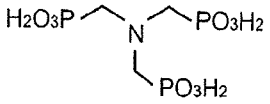
1,2-Ethanedisulfonic acid (EDSA)



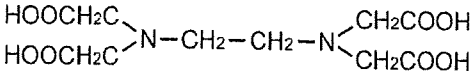
1,2-Ethanediphosphonic acid (EDPNA)



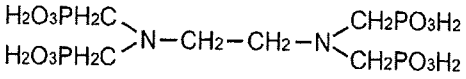
Nitrilotriacetic acid (NTA)



Nitrilotris(methanephosphonic acid) (NTMPNA)



Ethylenediaminetetraacetic acid (EDTA)



Ethylenediaminetetrakis(methanephosphonic acid) (EDTMPNA)

Figure 1. Structures and abbreviations of candidate acids in this study.

might be true for the cases of NTMPNA and EDTMPNA, from the comparison with NTA and EDTA. Multiplicity of the acidic group in one molecule dramatically improves the enhancement (from the data of MPNA, MDPNA, EDPNA and so on). The data for PA and the most promising candidate, MDPNA, indicate that MDPNA addition produces a 5-fold to 34-fold increase in signal enhancement whereas PA addition exhibits a 1-fold to 8.4-fold increase, compared with no addition of matrix additive (DHBA only). As a whole, it is apparent that the phosphonic acid moiety has an enhancing effect and that a molecule incorporating 2 to 4 of the phosphonic acid groups enhances the phosphopeptide signal in MALDI-MS. Among candidates incorporating multiple phosphonic acid groups, MDPNA is the most effective in enhancing the phosphopeptide signal in MALDI-MS.

In the negative mode, the data exhibit a similar tendency in that candidates incorporating 2 to 4 phosphonic acid groups are effective in enhancing the signal intensity of phosphopeptide ions by MALDI-MS, although the enhancing effect is not as high as in the positive mode. MDPNA is the most effective in enhancing the phosphopeptide signal, although the difference between PA and MDPNA becomes smaller.

Among the candidate compounds tested, MDPNA enhances the phosphopeptide signal by MALDI-MS the most in both positive and negative modes. Hence, MDPNA was used as an additive to DHBA for the subsequent experiments.

To optimize the concentration of the MDPNA solution, we tested several points of the concentrations ranging from 0.01% to 10%. At a concentration less than 1% or more than 5%, mass spectra indicated deteriorated signal-to-noise

DOI: 10.1002/rcm

RCM

Table 1. Phosphopeptides studied in this report*

Peptide ID	$[M+H]^+$	Sequence
1P	1438.6	MHRQEpTVDCLK-NH_2
2P	1880.2	AAKIQApSFRGHMARKK
3P	1702.8	TRDIpYETDYYRK
3PPP	1862.8	TRDIpYETDpYpYRK
OP1	2088.9	EVVGpSAEAGVDAASVSEEFR
OP2	2511.1	LPGFGDpSIEAQCGTSVNVHSSLR
OP3	2901.3	FDKLPGFGDpSIEAQCGTSVNVHSSLR
CP1	769.4	VNELpSK
CP2	1660.8	VPQLEIVPNpSAEER
CP3	1927.7	DIGpSEpSTEDQAMEDIK
CP4	1952.0	YKVPQLEIVPNpSAEER
CP5	2720.9	QMEAEpSIpSpSpSEEIVPNpSVEQK

*Commercially available phosphopeptides are denoted with 1P-3PPP, in which the number of P's indicates the degree of phosphorylation. Phosphopeptides from ovalbumin are presented with OP1-OP3, which are all monophosphorylated peptides. Phosphopeptides from α-casein are presented with CP1–CP5, of which CP-1, CP2 and CP4 are monophosphorylated, CP3 is diphosphorylated, and CP5 is pentaphosphorylated.

(S/N) ratio (data not shown). However, at a concentration in the 1–5% range, optimized effect was obtained in mass spectra. Hence we chose the lowest concentration in the range, 1%, for the matrix-additive solution.

Effect of MDPNA in eliminating alkali-metal adduct ions

Several additives such as phosphoric acid,[17] diammonium citrate,[16] etc., are reported to reduce or eliminate the signal of the alkali-metal adduct ion as well as to enhance phosphopeptide signals by MALDI-MS. Suppressing adduct formation is of great benefit for MALDI analysis. The signal enhancement of the peptide ion $[M+H]^+$ is supplemented by the reduction or elimination of alkali-metal ion adducts such as $[M+Na]^+$ and $[M+K]^+$. It was reported that these alkali metals may be captured by the additives, which is attributed to the ability of additives to form alkali-metal salts of acids such as carboxylic acid or phosphoric acid ('salting out').[17]

To test the effect of MDPNA, a peptide mixture solution, containing WAGGDASGE ($[M+H]^+$; 849.3), WAGGDApSGE ($[M+H]^+$; 929.3),GFETVPETG-NH_2 ($[M+H]^+$; 935.4), GFETVPEpTG-NH_2 ($[M+H]^+$; 1015.4), TSTEPQYQPGENL ($[M+H]^+$; 1463.6), TSTEPQpYQPGENL ($[M+H]^+$; 1543.6), with a concentration of 1.7 pmol/μL each was prepared.

Figure 2 depicts the elimination effect of alkali-metal adduct signal using MDPNA as an additive to the DHBA matrix. Adduct signals accompany these five peptide signals except for the signal of GFETVPEpTG-NH_2 (*m/z* 1014.4)

Table 2. Typical signal intensities as well as S/N ratios of the four phosphopeptides (1P, 2P, 3P, and 3PPP) with or without the addition of inorganic/organic acid as an additive to DHBA matrix. All experiments were replicated five times. The signal intensities are averaged of five replicated data, and are relative to those with no additive (DHBA only)

		1P		2P		3P		3PPP	
Comatrix	Ion mode	Signal	S/N	Signal	S/N	Signal	S/N	Signal	S/N
DHBA only	Positive	1.0	56	1.0	93	1.0	16	1.0	38
PA	Positive	1.2	62	3.1	243	8.4	113	1.5	48
PNA	Positive	3.1	42	6.1	200	4.6	84	1.5	18
MPNA	Positive	0.3	65	0.6	179	0.9	124	0.3	72
BPNA	Positive	—	—	—	—	0.1	5	—	—
MLA	Positive	0.2	21	0.3	62	0.3	9	0.1	8
SAA	Positive	1.1	62	1.7	149	6.3	141	1.3	61
PNAA	Positive	3.0	88	5.9	209	12.1	127	2.9	72
MDPNA	Positive	5.4	95	8.3	237	34.2	172	6.7	79
EHDPNA	Positive	1.3	82	2.7	201	10.5	177	2.1	70
SCA	Positive	0.8	28	1.1	65	1.1	11	0.6	14
EDSA	Positive	0.7	68	1.0	167	3.9	111	0.3	19
EDPNA	Positive	3.8	89	7.0	210	24.7	166	3.6	64
NTA	Positive	1.4	76	1.9	167	4.2	64	1.8	66
NTMPNA	Positive	4.2	83	6.9	211	21.6	125	6.2	71
EDTA	Positive	1.2	61	1.9	161	2.4	35	1.9	66
EDTMPNA	Positive	2.1	71	4.0	214	6.8	64	2.2	38
DHBA only	Negative	1.0	57	1.0	62	1.0	23	1.0	12
PA	Negative	1.0	39	3.7	86	7.2	118	1.7	13
PNA	Negative	0.6	25	1.6	68	3.0	52	1.2	10
MPNA	Negative	0.6	44	1.0	73	1.1	33	1.3	19
BPNA	Negative	—	—	—	—	0.1	1	—	—
MLA	Negative	0.1	7	0.2	10	0.2	4	0.1	1
SAA	Negative	0.1	7	0.1	6	0.9	27	—	—
PNAA	Negative	0.7	42	1.7	90	2.8	73	0.9	12
MDPNA	Negative	3.2	76	3.7	87	15.4	140	5.7	26
EHDPNA	Negative	1.8	74	2.2	86	8.8	134	2.2	17
SCA	Negative	0.3	13	0.4	17	0.4	6	0.5	4
EDSA	Negative	—	—	—	—	—	—	—	—
EDPNA	Negative	2.7	62	3.6	78	14.5	144	3.7	16
NTA	Negative	0.6	43	1.0	73	1.3	38	1.2	18
NTMPNA	Negative	2.1	86	2.6	81	8.8	136	4.4	35
EDTA	Negative	1.3	25	2.4	51	2.1	17	4.4	20
EDTMPNA	Negative	1.6	51	2.2	63	3.5	40	2.0	12

DOI: 10.1002/rcm

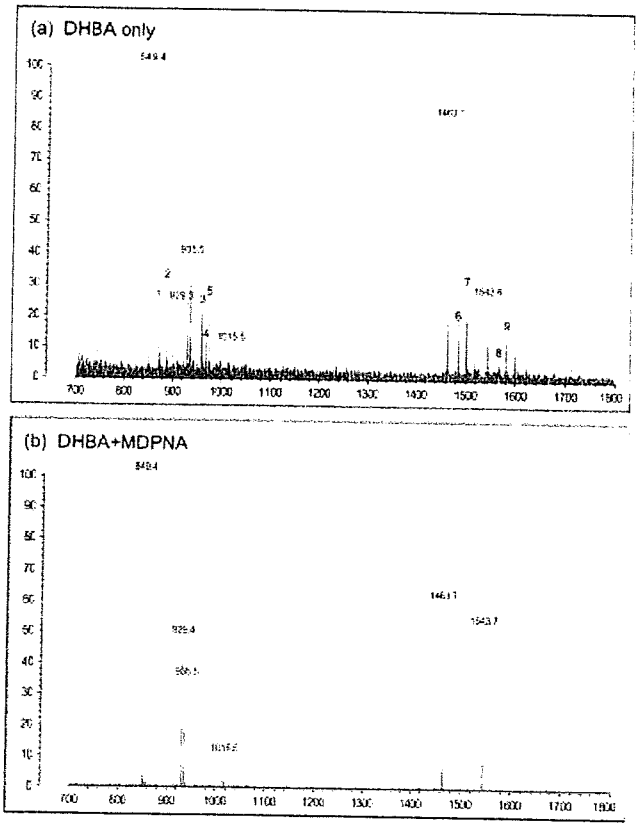


Figure 2. Effect of MDPNA addition in eliminating alkali-metal adduct ions. Intensive adduct formation is observed, except GFETVPEpTG-NH_2 (*m/z* 1014.4), with nothing added to the DHBA matrix (a). However, no adduct signal is observed when MDPNA is added to the DHBA matrix (b). The adduct signals (1–9) in (a) are as follows; 1. *m/z* 871.4 (848.3+Na), 2. *m/z* 887.4 (848.3+K), 3. *m/z* 957.5 (934.4+Na), 4. *m/z* 967.3 (928.3+Na), 5. *m/z* 973.4 (934.4+K), 6.*m/z* 1485.7 (1462.6+Na), 7.*m/z* 1501.6 (1462.6+K), 8. *m/z* 1565.6 (1542.6+Na), 9. *m/z* 1581.6 (1542.6+K).

without addition of MDPNA (Fig. 2(a)). The adduct signals indicated with numbers (1–9) in Fig. 2(a) completely disappeared after addition of MDPNA (Fig. 2(b)).

In this study, significant improvement of sample homogeneity was observed when adding MDPNA to DHBA (data not shown). The quality improvement of the spectrum, as can be seen in Figs. 2(a) and 2(b), may be partly attributed to the reduction of the inhomogeneity causing hot-spot formation.

Efficient detection of phosphopeptides using MDPNA

It was mentioned that fewer peptide ions were produced in negative MALDI-MS mode in the low *m/z* region.[17] In our study, we observed that peptide signals with or without phosphorylation sometimes escaped detection even in positive mode when using PA as an additive. However, this phenomenon was not observed with the use of MDPNA. Figure 3 graphs the spectra of a tryptic digest of α-casein as an example in the range of *m/z* 600–1000; arrows in the spectrum (Fig. 3(b)) indicate the non-detected signals when

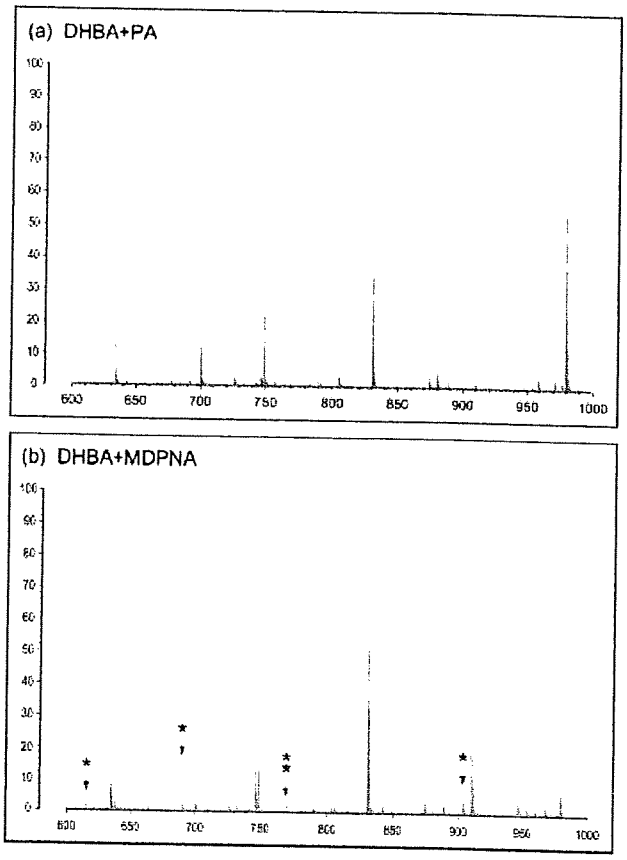


Figure 3. Efficient detectability of phosphopeptides using MDPNA as an additive to DHBA matrix. (a, b) MALDI mass spectra of α-casein digests. Signals that escaped detection using PA as an additive (a) are discernible when using MDPNA as an additive (b).

using PA (Fig. 3(a)). The phosphorylated peptide signal is indicated with a double asterisk; a single asterisk indicates the non-phosphorylated peptide signal.

Phosphopeptide mapping by MALDI-MS

To investigate the general applicability of MDPNA to phosphopeptide analysis by MALDI-MS, we attempted to analyze phosphopeptide fragments contained in a digest of a model phosphoprotein, ovalbumin, and a four-protein mixture (BSA, phosphorylase b, lysozyme, and α-casein).

1. Phosphopeptide mapping of ovalbumin

The tryptic digestion of ovalbumin generates three phosphopeptides (OP1, OP2, and OP3; see Table 1) as well as numerous non-phosphorylated peptides up to *m/z* 3000. Figure 4 presents six panels of mass spectra in three sample preparation conditions with addition of no acid (a, d), PA (b, e), and MDPNA(c, f) and with two sampling amount of ovalbumin digest (30 fmol/well for a,b, and c; and 3 fmol/well for d, e, and f). The phosphopeptide OP1 was observed in the three conditions employed at 30 fmol/well, though the signal from the phosphopeptide was rather low (a, b). The ion signals of OP2 and OP3 were, at 30 fmol/well sampling, not detected under the conditions of no acid (a) (DHBA only) and PA (b). However, these two signals as well as the OP1

RCM

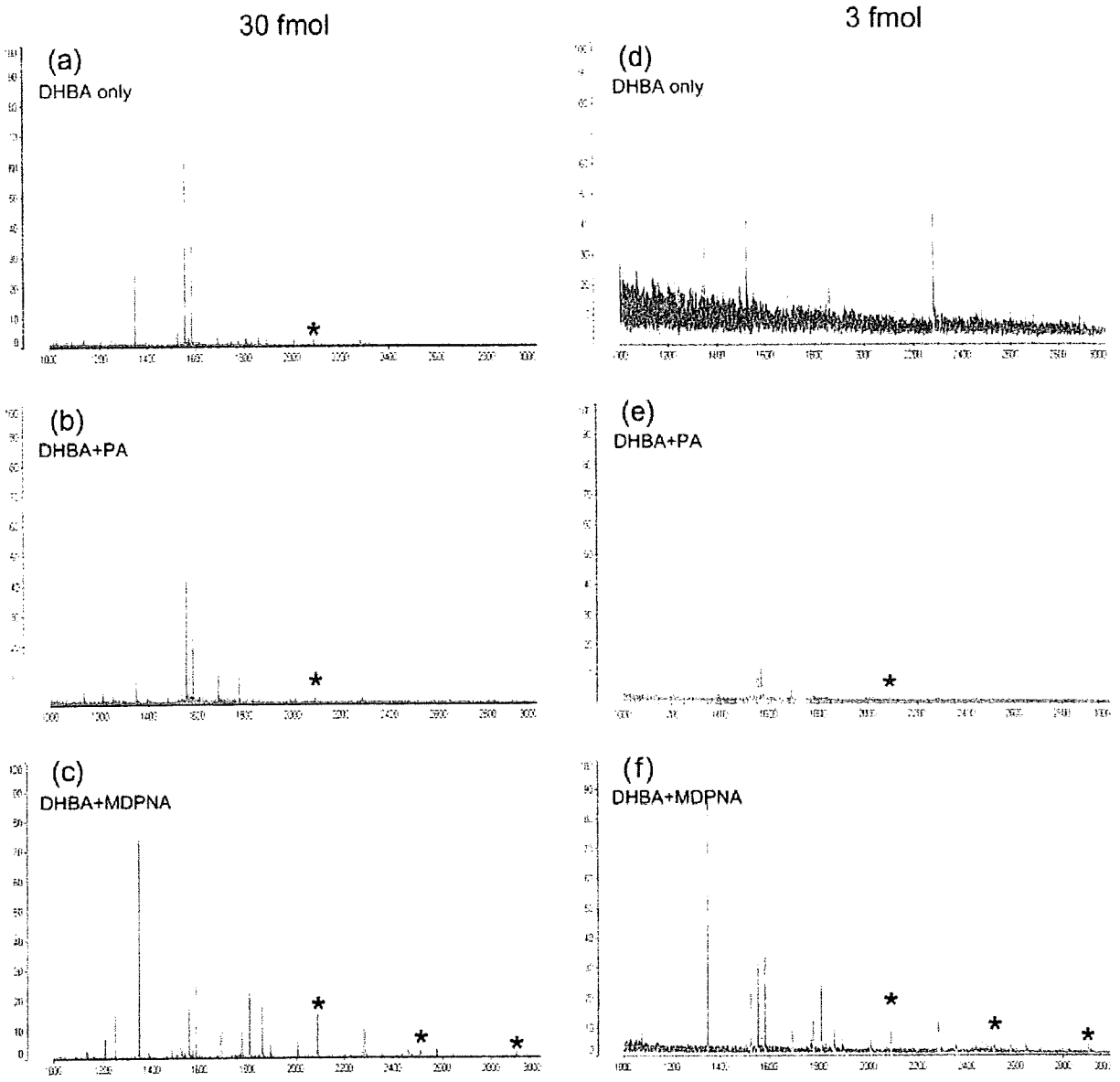


Figure 4. MALDI mass spectra of tryptic digest of ovalbumin. The tryptic digest was diluted with water to adjust the concentrations to 90 fmol/μL or 9 fmol/μL, from which a portion (0.3 μL) was applied to a target plate with or without an additive: no additive (a, d), PA as an additive (b, e), MDPNA as an additive (c, f). Signals with an asterisk represent phosphopeptides of OP1 ($[M+H]^+$: 2088.9), OP2 ($[M+H]^+$: 2511.1), and OP3 ($[M+H]^+$: 2901.3).

signal were clearly detected when MDPNA was used as a matrix additive (c). At 3 fmol/well, no signals from these phosphopeptides were detected with use of DHBA only (d), and one signal from the OP1 phosphopeptide was observed, although OP2 and OP3 escaped detection using PA (e). However, using MDPNA as an additive to DHBA, these three phosphopeptides were clearly observed even at 3 fmol/well sampling (f).

2. *Phosphopeptide mapping of a four-protein mixture*

We next tried a four-protein mixture containing BSA, phosphorylase b, lysozyme, and α-casein. Of the four proteins α-casein is a typical phosphoprotein, from which phosphopeptides generated after proteolysis were analyzed. For testing MDPNA as a matrix additive, two types of solution were prepared; (1) 2.1 ng/μL for BSA, phosphorylase b, lysozyme, and α-casein; (2) 2.1 ng/μL for BSA, phosphorylase b and lysozyme, and 0.21 ng/μL for α-casein. In the solution of (1), the concentration of α-casein was 90 fmol/μL and in (2), 9 fmol/μL. For MALDI-MS analysis 0.3 μL of sample solution was applied onto a MALDI target plate. Hence, the sampling amounts of α-casein were 30 fmol/well and 3 fmol/well, respectively. Five phosphorylated peptides (CP1, CP2, CP3, CP4 and CP5; see Table 1) were generated after tryptic digestion along with non-phosphopeptides up to *m/z* 3000. Figure 5 presents six panels of mass spectra in three sample preparation conditions, as in Fig. 4, with addition of no additive (a, d), PA (b, e), and MDPNA (c, f) for the two model solutions. At 30 fmol/well of α-casein, all of five phosphopeptides were observed when using MDPNA as an additive (c); however, CP1 was not detected when using PA (b), and CP1 and CP3 were missing

DOI: 10.1002/rcm

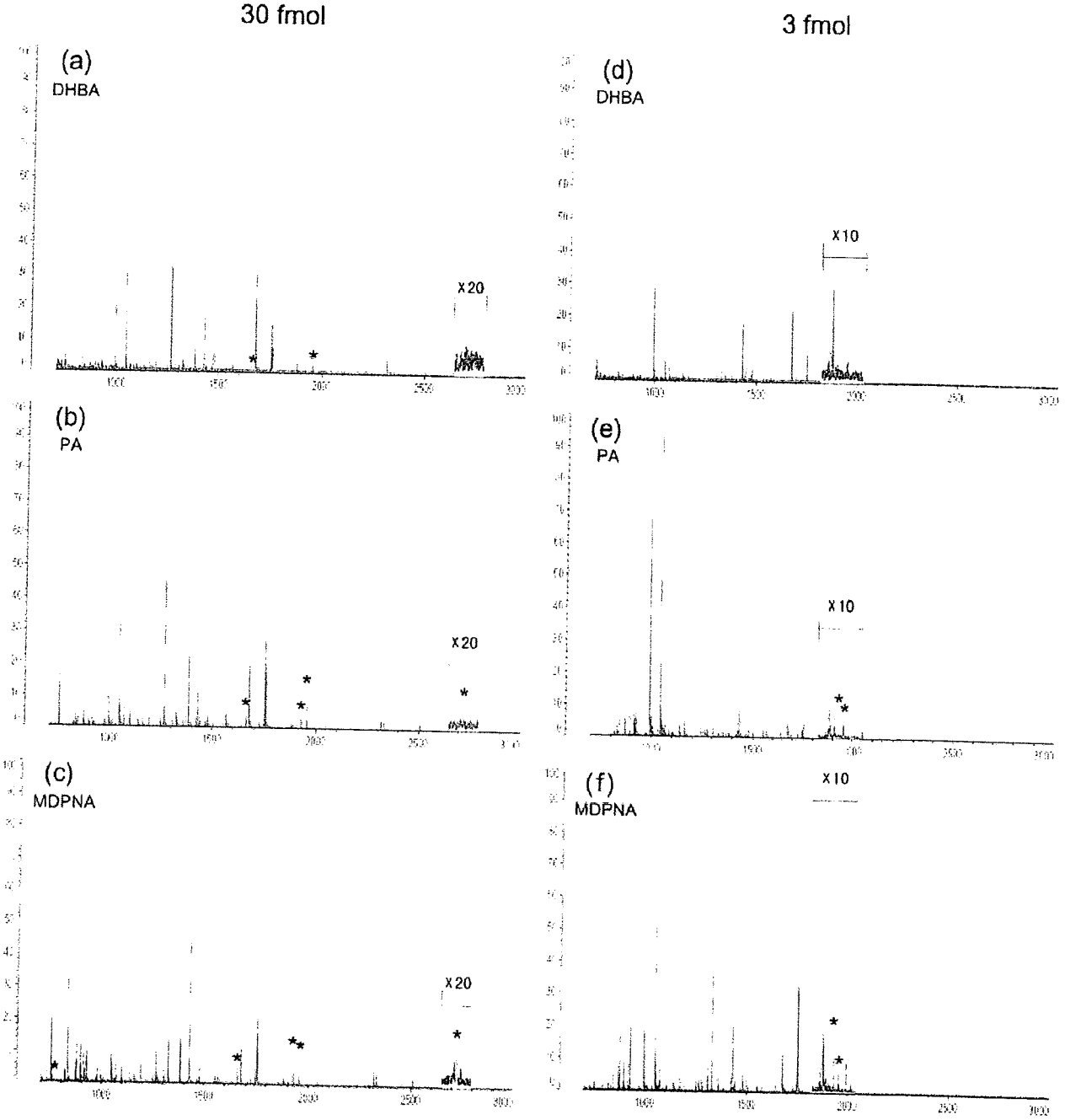


Figure 5. MALDI mass spectra of tryptic digest of the four-protein mixture. Two model mixtures were prepared that contained α-casein as a component at 90 fmol/μL and 9 fmol/μL, from which a portion (0.3 μL) was applied to a target plate with or without an additive: no additive (a, d), PA as an additive (b, e), MDPNA as an additive (c, f). Signals with an asterisk represent phosphopeptides of CP1 ($[M+H]^+$: 769.4), CP2 ($[M+H]^+$: 1660.8), CP3 ($[M+H]^+$: 1927.7), CP4 ($[M+H]^+$: 1952.0) and CP5 ($[M+H]^+$: 2720.9).

without an additive (a). At 3 fmol/well of α-casein, no signals of these five phosphorylated peptides were detected when using DHBA only (d). CP1, CP2 and CP5 were not observed even with PA or MDPNA addition. Though CP3 and CP4 were observed in both (e) and (f), the S/N ratios of the two signals were better with MDPNA addition (f). In this experiment using a four-protein mixture, the enhancement efficiency of MDPNA was clearly demonstrated in comparison with no additive (DHBA only).

These results indicated that adding MDPNA to DHBA matrix significantly enhances detection of ovalbumin and α-casein phosphopeptides by MALDI-MS.

When DHBA is used as a matrix, an intrinsic problem is pronounced hot-spot formation due to inhomogeneous sample preparation, which causes prolonged measurement times and is unfavorable for automated data acquisition. The use of MDPNA and other compounds incorporating multiple phosphonic acid moieties improves sample

DOI: 10.1002/rcm

research articles Journal of proteome research

Novel Breast Cancer Biomarkers Identified by Integrative Proteomic and Gene Expression Mapping

Keli Ou,[†,∥] Kun Yu,[‡] Djohan Kesuma,[†,∥] Michelle Hooi,[†] Ning Huang,[†] Wei Chen,[†] Suet Ying Lee,[†] Xin Pei Goh,[†] Lay keng Tan,[†] Jia Liu,[†] Sou Yen Soon,[†] Suhaimi Bin Abdul Rashid,[#] Thomas C. Putti,[#] Hiroyuki Jikuya,[†,∥] Tetsuo Ichikawa,[†,∥] Osamu Nishimura,[§] Manuel Salto-Tellez,[#] and Patrick Tan*[,†,∥,⊥,¶]

Agenica Research Pte Ltd., National Cancer Centre of Singarope, and Genome Institute of Singapore, 11 Hospital Drive, Singapore 169610, Shimadzu (Asia Pacific), 16 Science Park Drive, Singapore 118227, Shimadzu Corporation, Kyoto, Japan 604-8511, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, and Duke-NUS Graduate Medical School, 2 Jalan Bukit Merah, Singapore 169547

Received December 6, 2007

Proteomic and transcriptomic platforms both play important roles in cancer research, with differing strengths and limitations. Here, we describe a proteo-transcriptomic integrative strategy for discovering novel cancer biomarkers, combining the direct visualization of differentially expressed proteins with the high-throughput scale of gene expression profiling. Using breast cancer as a case example, we generated comprehensive two-dimensional electrophoresis (2DE)/mass spectrometry (MS) proteomic maps of cancer (MCF-7 and HCC-38) and control (CCD-1059Sk) cell lines, identifying 1724 expressed protein spots representing 484 different protein species. The differentially expressed cell-line proteins were then mapped to mRNA transcript databases of cancer cell lines and primary breast tumors to identify candidate biomarkers that were concordantly expressed at the gene expression level. Of the top nine selected biomarker candidates, we reidentified ANX1, a protein previously reported to be differentially expressed in breast cancers and normal tissues, and validated three other novel candidates, CRAB, 6PGL, and CAZ2, as differentially expressed proteins by immunohistochemistry on breast tissue microarrays. In total, close to half (4/9) of our protein biomarker candidates were successfully validated. Our study thus illustrates how the systematic integration of proteomic and transcriptomic data from both cell line and primary tissue samples can prove advantageous for accelerating cancer biomarker discovery.

Keywords: breast cancer • proteomics • transcriptomics • bioinformatics • integrative genomics

Introduction

Breast cancer is a major worldwide cause of morbidity and mortality in females.[1,2] Patients diagnosed with early stage breast cancer have been shown to experience significantly improved survival compared to late stage patients,[3] making the identification of molecular biomarkers to facilitate early detection and screening an important goal.[3–6] Currently, the majority of biomarkers in clinical use are either protein or antibody based, due to their high sensitivity and specificity, reproducibility, and robustness to different sample types including paraffin-embedded archival samples, frozen samples, and body fluids. The need for novel cancer biomarkers has contributed significantly to the increasing number of proteomic studies analyzing cancer cells and tumors of various types.[7–9] A significant limitation, however, is that the proteomic analysis of primary tumor samples is widely recognized in the field as highly technically challenging. First, primary tumors are often small in quantity and often contain several distinct cellular populations including tumor cells, stroma, immune cells, and blood vessels, which can often lead to significant variability in protein profiles between different individuals.[10] Second, protein contributions from different compartments, particularly blood, can often overwhelm tumor-intrinsic signals due to the presence of highly abundant plasma proteins such as IgG and serum albumin. Third, it is still prohibitive at most centers to proteomically profile a statistically meaningful number of primary tissue samples in terms of cost, labor, and time. While tumor cell lines are a convenient alternative where large amounts of genetically homogeneous material can be generated, the extension of proteomic cell line studies to the primary tumor setting is often confounded by differences associated with *in vitro* and *in vivo* growth.[10]

* To whom correspondence should be addressed. Dr. Patrick Tan, Duke-NUS Graduate Medical School, 2 Jalan Bukit Merah, Singapore 169547. Tel: (+65) 64368345. Fax: (+65) 62265694. E-mail: gmstanp@nus.edu.sg.
† Agenica Research Pte Ltd.
∥ Shimadzu (Asia Pacific) Pte Ltd.
‡ National Cancer Centre of Singapore.
# National University of Singapore.
§ Shimadzu Corp.
⊥ Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672.
¶ Duke-NUS Graduate Medical School.

Published on Web 03/05/2008

10.1021/pr700820g CCC: $40.75

In parallel to proteomics, gene expression profiling has also played an important role in the identification of biomarker signatures for patient prognosis and treatment response prediction. Compared to proteomics, expression profiling platforms offer relatively higher sensitivity and much higher throughput, as hundreds of samples can be profiled in a relatively short amount of time.[11,12] The major limitation of gene expression profiling, in the context of protein biomarker discovery, is that there is often a less than perfect correlation between gene and protein expression,[13,14] and it has been shown that variations between a gene's mRNA level and its protein abundance can be as high as 30-fold.[15] Furthermore, important post-translational modification events, such as protein phosphorylation or cleaveage events, are obviously not captured in gene expression data.[16] To improve the efficiency of biomarker discovery, it would thus be useful to employ strategies combining the advantages of both expression profiling and proteomics, while minimizing the limitations of both platforms.

To address these problems, we present in this study a integrated proteo-transcriptomic strategy for biomarker discovery combining both the direct protein visualization capacity of proteomics with the high-throughput advantages of gene expression profiling. Using a two-dimensional electrophoresis (2DE)/mass spectrometry (MS) proteomics platform, total cell lysates of breast cancer (e.g., MCF-7 and HCC-38) and control cell lines (e.g., CCD-1059Sk) were analyzed to identify >1500 protein spots. We then used mRNA transcript databases of breast cell lines and primary tumors to prioritize these differentially expressed proteins by their consistency of gene expression differences between tumors and normal tissues. Using this strategy, we identified nine potential breast cancer biomarker candidates, and independently validated four of them (ANX1, CRAB, 6PGL, CAZ2) by immunohistochemistry on breast cancer tissue microarrays. In summary, our results demonstrate how the integration of proteomic and gene expression data may prove useful in accelerating biomarker discovery.

Materials and Methods

Cell Lines and Primary Tumors. The ER positive human breast carcinoma cell line MCF-7 (ATCC: HTB-22) and the nonmalignant control cell line CCD-1059Sk (ATCC: CRL-2072) were cultured in minimum essential medium (Eagle) and Dulbecco's modified eagle medium, respectively, supplemented with 2 mM L-glutamine and Earle's BSS adjusted to contain 1.5 g/L sodium bicarbonate, 0.1 mM nonessential amino acids, 1 mM sodium pyruvate, and 10% fetal bovine serum. The ER negative human breast carcinoma cell line HCC-38 (ATCC: CRL-2314) was cultured in RPMI 1640 medium with 2 mM L-glutamine adjusted to contain 1.5 g/L sodium bicarbonate, 4.5 g/L glucose, 10 mM HEPES, 1 mM sodium pyruvate, and 10% fetal bovine serum. The cell lines were cultured in a humidified incubator with 5% CO_2 at 37 °C. Cells were grown to confluence and then incubated with three changes of serum-free medium over 48 h. Cell line samples were prepared by washing cells with ice-cold PBS and trypsinized for 5 min to detach cells. The suspension was transferred to a 15 mL tube and the cells were centrifuged at 1000*g* for 5 min. The supernatant was discarded and 1–2 mL of PBS was added to resuspend the cells prior to cell counting. About 5×10^6 cells were aliquoted into each 1.5 mL Eppendorf tube and centrifuged at 380g for 10 min. The cell pellets were snap-frozen in liquid nitrogen and stored at −80 °C.

Primary human breast tissues used for protein and mRNA analyses were obtained from the Singapore National Cancer Centre Tissue Repository. Detailed descriptions of sample collection, archiving, and histological assessment of tumors including techniques and scoring parameters have been previously reported.[12]

Preparation of Protein Extracts and 2DE Separation. The same batch of cell line lysates was split for both protein and mRNA profiling analyses. For each cell line, we profiled 3 independently grown batches. For proteomic analysis, each cell line batch was also repeated across 3 to 4 technical replicates. For the first two batches, each of the three cell lines had 3 replicate gels at two pH levels (pH 4–7 and pH 6–9), while for the third batch, each cell line had 3 replicate gels at pH 4–7 and 4 replicate gels at pH 6–9. In total, 57 gels were run. Protein extraction and 2DE separation was carried out as previously described.[17] Briefly, harvested cells were disrupted with a cocktail of 7 M urea (Bio-Rad), 2 M thiourea (Fluka Chemie), 4% (w/v) CHAPS (GE HealthCare), 1 mM PMSF (Sigma), 50 μg/mL DNAse (Boehringer Mannheim), 50 μg/mL RNAse (Boehringer Mannheim), and protein inhibitor cocktail (Sigma, 1 mL/108 cells). The resulting cell lysate was sonicated with a probe sonicator (Misonix, Inc., NY) and centrifuged at 40 000 rpm for 20 min to remove cellular debris. The extracted proteins were cleaned up using a 2D-sample clean up kit (Bio-Rad). Protein quantification was determined using the Bradford method.

The first dimensional protein separation was performed on the IPGphor IEF system (GE HealthCare) using IPG strips (24 cm, pH 4–7 and pH 6–9, respectively, GE HealthCare) that had been rehydrated with 450 μL of rehydration buffer (7 M urea, 2 M thiourea, 4% CHAPS, 0.5% IPG buffer, and 20 mM DTT) for 8 h. For the IPG 6–9 strips, the destreak rehydration buffer (GE HealthCare) was used as rehydration solution. Protein lysates of 200 μg in 80 μL of lysis buffer containing 60 mM DTT and 0.5% IPG buffer were applied through cup loading anodically. The strips were then focused using the following electrical conditions at 20 °C: 300 V for 2 h; 300−1000 V for 3 h; from 1000 to 8000 V for 4 h; and 8000 V until a total of 72 000 Vh was reached. After IEF focusing, the IPG strips were equilibrated in 5 mL of equilibration buffer (pH 6.8) containing 1% DTT and rocked for 15 min. The strips were then soaked in 5 mL of equilibration buffer (pH 8.8) containing 2.5% IAA and rocked for 15 min. After equilibration, the strips were transferred onto 10% isocratic polyacrylamide slab gels (24 cm × 20 cm × 0.75 mm). The IPG gels were sealed with 0.5% (w/v) agarose in running buffer (25 mM Tris-HCl, 192 mM glycine, and 0.1% SDS, pH 8.3) on top. The second dimensional separation was performed with the Dodeca Cell system (Bio-Rad). The SDS-PAGE was carried out at 17 °C at a constant voltage of 200 V until the dye front reached the bottom of the gels. The gels were fixed and labeled with Deep Purple fluorescent dye (GE HealthCare) according to the manufacturer's instructions. The stained gels were digitized using a Molecular Imager FX system (Bio-Rad) and image analysis was carried out with the PDQuest 7.3 image analysis software (Bio-Rad). All data points of gel spot stain intensity were means of 3 or 4 analytical replicates.

MS Identification of Proteins. For protein identification, fluorescence stained gel spots were semiautomatically excised, washed, tryptic-digested, cleaned, and spotted onto the MS sample plate using the Shimadzu Xcise robotic system (Shimadzu, Kyoto, Japan). The excised gel spots were 1.6 mm in diameter and were washed with 50 mM NH_4HCO_3, pH 8.8,

Table 1. Comprehensive Proteome Analysis on Three Breast Cell Lines (CCD-1059Sk, MCF-7, and HCC-38) Using pH 4–7 and pH 6–9 IEF, Respectively[a]

	cell lines	no. of spots cut	no. of spots identified	no. of unique proteins identified %
pH 6–9	CCD-1059Sk	329	196	60
	MCF-7	323	217	67
	HCC-38	188	135	72
pH 4–7	CCD-1059Sk	440	315	72
	MCF-7	624	428	69
	HCC-38	644	433	67
	Total	2548	1724	67

[a] The proteins were identified by peptide mass fingerprinting using MALDI-TOF MS.

containing 50% acetonitrile (ACN) for 10 min. After dehydration with 100% ACN, the gel pieces were rehydrated with 30 μL of 50 mM NH_4HCO_3, pH 8.5, containing 3.3 μg/mL trypsin (Promega, Madison, WI) and incubated at 30 °C overnight. The samples were then cleaned and concentrated using μC18 ZipTips (Millipore). Finally, the peptide mixtures were eluted with 1.5 μL of 50% ACN/0.5% TFA onto a MALDI sample plate and mixed with 1 μL of matrices containing 5 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA) plus 5 mg/mL 2,5-dihydroxybenzoic acid (DHB) in 50% ACN/0.5% TFA prior to MS analysis. MALDI-TOF MS analysis was performed using the AXIMA-CFR plus mass spectrometer (Shimadzu Biotech, Manchester, U.K.). The operating conditions were as follow: nitrogen laser (337 nm); reflectron mode; detection of positive ions. The acceleration potential was set to 35 kV using a gridless-type electrode. MALDI-TOF MS spectra were acquired in the auto experiment mode, from m/z 800 to 3000, and internally calibrated with two trypsin autolysis peaks (m/z 842.51 and 2211.10). Peak lists from PMF spectra were used to automatically search against the UniProt or NCBInr human protein databases using the Kompact program (Shimadzu Biotech, Manchester, U.K.) through an in-house Mascot server V2.1.04 (Matrix Science, U.K.). To minimize the chance of false-positive protein identification, the database search results were manually validated based on the following criteria: (1) 0.1 Da or better mass accuracy; (2) most of the major peaks in the spectrum were matched to the tryptic peptide list of the protein; (3) MW and pI of the identified protein should match the estimated values from 2DE image analysis. In our hands, about 67% of the excised protein spots were positively identified (Table 1). Unidentified protein spots were excluded from subsequent analysis.

Gene Expression Profiling. Total RNAs were extracted from human tissue or cell line samples with Trizol reagent (Invitrogen, CA) and hybridized onto Affymetrix U133_plus GeneChips following the manufacturer's standard protocols. Raw scanned files were stored in a central database and quality controlled using GeneData Refiner software (GeneData, Basel, Switzerland). The refined profiles were mean-normalized to an intensity value of 500. Detailed description of the profiles have been previously reported.[12] The primary tumor gene expression data is available from the GEO database under accession number GSE2294.

Statistical Integration of Proteomics and Gene Expression. For enrichment analysis, protein IDs identified by MS analysis were matched to their corresponding U133_plus probes using Swiss-Prot IDs (www.affymetrix.com). To check the concordance between mRNA and protein levels for a set of genes, we applied Gene Set Enrichment Analysis (GSEA), a computational method that determines whether an a priori defined set of genes shows statistically significant and concordant differences between two biological states.[18] Briefly, the gene expression profiles of the 3 cell lines were filtered to retain genes exhibiting >60% of valid values across the data set. A "valid" measurement was defined by a present/absent call generated by the MAS5 statistical algorithm provided by Affymetrix. Three replicate hybridizations were performed per cell line. The gene list was ranked by the level of gene expression differences (signal-to-noise ratio) between MCF-7 versus CCD-1059Sk and between HCC-38 versus CCD-1059Sk, respectively. The significantly differentially expressed protein spots selected from the 2DE gel image comparisons were then mapped to their corresponding mRNAs and used as a test group against the ranked gene lists for GSEA analysis. The significance of the GSEA was determined by a sample-based permutation test where the sample labels were randomly shuffled. Concordance of gene and protein expression was defined as entities having the same fold change direction (i.e., >1 or <1), that is, the mRNAs were selected if they had the same fold change direction as the protein result. No p-value cutoffs were implemented.

Gene Rank Analysis. Candidate biomarkers were selected by first excluding candidates showing discordant trends between protein and mRNA levels. The concordantly expressed biomarkers were further stratified using a Support Vector Machine (SVM) algorithm.[19] Briefly, a cohort of 39 breast tissue samples (7 normal and 32 tumors) was used as the training data set. An SVM algorithm with a linear kernel and a penalty value of 10 was used for cross-validation. After 100 cross-validation runs, the biomarkers were ranked based on their individual contributions to the normal versus tumor classification accuracy. The final set of 9 biomarkers was then further validated against an independent cohort of 5 normal and 31 tumor samples.

Tissue Microarray and Immunohistochemical Analysis. Potential biomarker candidates were validated using a breast cancer tissue microarray (TMA) composed of tumor (98) or normal (98) breast tissues.[20] The TMA was constructed using a tissue arraying instrument (ATA100, Advanced Tissue Arrayer, Chemicon, CA) and 1 mm-diameter tissue cylinders. All tumor samples were arrayed in duplicate, and histologically verified for sampling accuracy and adequacy. Consecutive sections were prepared on charged polylysine-coated slides for immunohistochemical (IHC) analysis. This approach ensures maximum reliability in the analysis of biomarkers in breast cancer[21] and has been used successfully applied in the validation of biomarker in previous breast cancer related studies.[22] This work was performed in the context of a TMA program supported by the Institutional Review Board of the National University of Singapore (NUS-IRB 05–017).

For TMA immunohistochemistry, individual TMAs were deparaffinized for 20 min at 60 °C, xylene-rinsed, and rehydrated in a series of alcohol/water rinses. The primary antibodies against the potential biomarkers were obtained either from commercial sources or custom-made by BioGene (Berlin, Germany). To ensure accurate optimization for each antibody, we analyzed several concentrations and 4 antigen retrieval methods per concentration. Antigen retrieval was performed by boiling in one of four different buffers (citrate buffer, Tris-EDTA, DAKP, pH 6.0, or DAKO, pH 9.0) in a pressure cooker or microwave. The custom-made primary antibodies were

further optimized using paraffin-embedded cell line blocks as positive controls, following a cell line processing protocol akin to routine clinical samples. Briefly, MCF-7 breast cancer cells were cultured in 75 mL flasks, and upon confluence, the cells were trypsinized and spun down. Immediately, the cell pellets were incubated with formalin for 20 min and paraffin-embedded. To evaluate antigen expression, protein subcellular location information was obtained from either: (1) datasheets attached to the commercially available antibodies and prediction using bioinformatic analysis,[23] or (2) in-house Western blotting experiment on MCF-7 cells subcellularly fractionated using the Calbiochem's Subcellular Proteome Extraction Kit (Merck). Step (2) was essential in correlating the antibody expression on positive control tissues to the tissue microarrays, to ensure appropriate biological interpretation. Each primary antibody was applied for 30 min at room temperature, and the detection of bound antibody was accomplished with the DAKO Envision Kit (DAKO, CA). The slides were dehydrated in a series of solutions: 70% ethanol, 95% ethanol, 4× 100% ethanol, and 2× xylene and mounted for visualization. Further information regarding the IHC analysis, that is, antigen retrieval method selection, source of primary antibodies, primary antibody titering, and protein subcellular localization, is presented in Table 2. To ensure homogeneity on the scoring methods, all TMA slides were examined and scored by a single pathologist (MST) using the scoring system of 0 to 3. Statistical associations of immunohistochemical staining patterns between tumors and normals were determined using a Fisher test.

Results

Proteomic Analysis of Breast Cancer Cell Lines. Figure 1 illustrates the schematic of our integrated proteo-transcriptomic approach for breast cancer biomarker discovery, demonstrating the effect of biomarker shortlisting. We prepared cell lysates from two breast cancer cell lines (MCF-7 and HCC-38) and a normal control cell line (CCD-1059Sk) and separated the proteins by 2DE using narrow-ranged pH IPG strips for pH ranges 4–7 or pH 6–9, on 10% isocratic SDS-PAGE slab gels. To minimize gel-to-gel variation, we used a Bio-Rad Dodeca Cell system, running up to 12 24-cm slab gels simultaneously. Deep Purple dye was used for postgel fluorescence labeling, and quantitative gel image analysis was performed using the PDQuest image analysis software (Supplementary Figures S1 and S2 in Supporting Information). In total, three independently grown batches of cell line samples were subjected to proteome analysis, with a gel spot concordance between the three independent batches of close to 90% (data not shown). The protein sets described in Table 1 are based on the batch 3 samples. We performed large-scale protein identification, excising a total of 2548 protein spots from the 2DE gels of the 3 cell lines using the Xcise robotic gel processing system. To increase the protein identification rate, at least 3 spots of the same protein from replicate gels were pooled and processed for MS analysis. A total of 1724 protein spots representing 484 different protein species were positively identified by PMF, a success rate of approximately 67%. Table 1 summarizes the number of protein spots identified in each sample. The annotated 2DE reference maps of the 3 cell line proteins are available in Supplementary Figures S3–S8, and the complete list of identified proteins in Supplementary Table S1 in Supporting Information. Some gel spots contained more than one protein, perhaps representing comigrating proteins with similar primary structures or molecular weights. For example, spot HCC38_pH4–7_087 was identified as tubulin alpha-1 chain (TBA1_HUMAN) or tubulin alpha-6 chain (TBA6_HUMAN), both proteins share some identical peptides, while spot HCC38_pH4–7_476 contained two different proteins of Annexin A5 (ANX5_HUMAN) and Cytokeratin 10 (K1CJ_HUMAN) (Supplementary Table S1 in Supporting Information) which have similar molecular weights. The proteins, including isoforms and variants, were then grouped into 13 categories based on their functions (Figure 2), including cytoskeleton and associated proteins, metabolic enzymes, molecular chaperones/heat shock proteins, membrane-associated proteins with multiple activities, calcium-binding proteins with EF-hand domain, proteins with binding functions, protein biosynthesis, nucleotide biosynthesis, cell growth and proliferation regulators, protein degradation and detoxification, and redox proteins.[24] Proteins with fragments or uncertain functions were placed under the category of 'others'. We found that the majority of the proteins were metabolic enzymes (35%), followed by cytoskeleton and associated proteins (12%) and molecular chaperones/heat shock proteins (10%). These functional classifications, however, should be treated with caution as they are subject to change, and some proteins may have multiple functions. We also found that over 50% of the protein species identified in this study contained isoforms or variants. For instance, gamma actin (ACTG_HUMAN) contained as many as 27 isoforms (Supplementary Table S1 in Supporting Information). The biological significance of isoforms is well-known[16] however, they are not our focus in the current work.

To identify differentially expressed proteins, we applied the Students t test on the protein spot stain intensity data to compare the MCF-7 and CCD-1059Sk protein profiles. Here, the replicates used in this comparison were all technical replicates to control for technical variations associated with the proteomic platform (e.g., gel casting, gel running, staining, etc.). There was no mixing of biological and technical replicates. For the pH 4–7 gels, we selected the top 200 spots ranked by their p-values. The maximum p-value for these spots was 0.0035 (i.e., all p-values < 0.0035). of these 200 spots, 108 had valid Swiss-Prot IDs and annotations. Similarly, for the pH 6–9 gels, we selected the top 200 spots, which were all associated with p-values < 0.0014. Of these, 82 spots had valid Swiss-Prot ID and annotations. Collectively, the 190 spots (108 + 82) corresponded to 64 unique proteins from the pH 4–7 gels and 57 unique proteins from the pH 6–9 gels, with 10 proteins overlapping between the two pH ranges. By combining data from the two pH ranges, we derived a final list of 78 overexpressed and 32 underexpressed proteins in MCF-7 relative to CCD-1059Sk. We then performed the same procedure to compare the HCC-38 and CCD-1059Sk protein profiles. Selecting the top 200 differently displayed protein spots ($p < 0.0067$ for pH 4–7 and $p < 0.00094$ for pH 6–9), we identified 62 differentially expressed proteins from the pH 4–7 gels (62 unique proteins) and 60 differentially expressed proteins from the pH 6–9 gels. By combining data from the pH ranges, we derived a final list of 81 overexpressed and 31 underexpressed proteins in HCC-38 cells compared to controls. A total of 56 proteins were found to overlap between the MCF-7 (110; Table S2 in Supporting Information) and HCC-38 (112; Table S3 in Supporting Information) protein comparisons. Of these, 35 proteins were commonly overexpressed when compared to CCD-1059Sk; 18 proteins were commonly underexpressed, and 3 proteins were discordant. A complete list of the differentially