Potential biological insights revealed by an integrated assessment of proteomic and transcriptomic data in human colorectal cancer

ICHIRO TAKEMASA[1*], NOBUYOSHI KITTAKA[1*], TOSHIKI HITORA[1], MAKOTO WATANABE[2], EI-ICHI MATSUO[2], TSUNEKAZU MIZUSHIMA[1], MASATAKA IKEDA[1], HIROHUMI YAMAMOTO[1], MITSUGU SEKIMOTO[1], OSAMU NISHIMURA[2], YUICHIRO DOKI[1] and MASAKI MORI[1]

[1]Department of Gastroenterological Surgery, Graduate School of Medicine, Osaka University; [2]Division of Disease Proteomics, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, Japan

Received July 20, 2011; Accepted September 1, 2011

DOI: 10.3892/ijo.2011.1244

Abstract. In the post-genomic era, the main aim of cancer research is organizing the large amount of data on gene expression and protein abundance into a meaningful biological context. Performing integrated analysis of genomic and proteomic data sets is a challenging task. To comprehensively assess the correlation between mRNA and protein expression, we focused on the gene set enrichment analysis, a recently described powerful analytical method. When the differentially expressed proteins in 12 colorectal cancer tissue samples were considered a collective set, they exhibited significant concordance with primary tumor gene expression data in 180 colorectal cancer tissue samples. We found that 53 upregulated proteins were significantly enriched in genes exhibiting elevated gene expression levels (P<0.001, ES=0.53), indicating a positive correlation between the proteomic and transcriptomic data. Similarly, 44 downregulated proteins were significantly enriched in genes exhibiting elevated gene expression levels (P<0.001, ES -0.65). Moreover, we applied gene set enrichment analysis to identify functional genetic pathways in CRC. A relatively large number of upregulated proteins were related to the two principal pathways; ECM receptor interaction was related to heparan sulfate proteoglycan 2 and vitronectin, and ribosome to RPL13, RPL27A, RPL4, RPS18, and RPS29. In conclusion, the integrated understanding of both genomic and proteomic data sets can lead to a better understanding of functional inference at the physiological level and potential molecular targets in clinical settings.

Introduction

The advent of genomic and proteomic technologies for the analysis of human tumor samples has now added an additional source of information to clinical cancer research (1). Genomic technologies have enabled rapid and sensitive screening for global and specific changes in gene expression that occur in hundreds of samples (2). Proteomic techniques based on mass-spectroscopy have enabled us to characterize and quantify thousands of proteins, which are closely related to the phenotype of an organism (3). Several previous studies have individually analyzed proteomic and transcriptomic data. Understanding the relevance and consistency of each data set individually is difficult and time-consuming. A major task in the post-genomic era is organizing this large amount of data on gene expression and protein abundance into a meaningful biological context.

We previously reported individual studies of mRNA and protein expression profile using DNA microarray and proteomics in colorectal cancer (CRC) (4,5). We believe that these studies should be expanded into a global integrated analysis of mRNA and protein expression profiles. However, a major limitation is that a less than perfect correlation often exists between mRNA and protein expression (6,7). Variations between the mRNA level of a gene and its corresponding protein abundance can be as high as 30-fold (8). Potential biological reasons for the lack of correlation between mRNA and protein expression levels are: i) translational regulation, ii) differences in protein *in vivo* half-lives, and iii) differences with respect to the experimental platforms (9,10). In addition, possible reasons for this poor correlation may include methodological constraints that might affect the comparison of mRNA and protein levels. Previous studies have reported that the Pearson correlation coefficient and Spearman rank coefficient for these data ranged from 0.47 to 0.76 in bacterial and mammalian cells (11-13).

To comprehensively assess the correlation between mRNA and protein expression in CRC, in this study, we focused on a recently described powerful analytical method known as the gene set enrichment analysis (GSEA) (14). GSEA determines

Correspondence to: Dr Ichiro Takemasa, Department of Gastroenterological Surgery, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka E-2, Suita 565-0871, Osaka, Japan
E-mail: itakemasa@gesurg.med.osaka-u.ac.jp

*Contributed equally

Abbreviations: CRC, colorectal cancer; GSEA, gene set enrichment analysis; NBS, 2-nitrobenzenesulfenyl

Key words: transcriptomics, proteomics, correlation analysis, gene set enrichment analysis, colorectal cancer

whether members of a biological motif set *(S)* tend to occur toward the top (or bottom) of the gene list *(L)* by testing the coordinated over- or under-expression of gene sets using the Kolmogorov-Smirnov test over a weighted summation (14). We hypothesized that the GSEA approach, which was not controlled by different dynamic range, might be more suitable to evaluate the correlation between mRNA and protein levels than the traditional correlation coefficient analysis. Of note, GSEA revealed a relationship that had not been previously evaluated by traditional analysis.

We sought to identify potential biological regulatory pathways in humans in large sample size (180 CRCs data) using this GSEA application that is also known as the global analytical approach to evaluate regulatory gene sets associated with several biological mechanisms of cancer. We examined the relevance of these pathways in transcriptomic and proteomic analysis of upregulated proteins. Of note, several upregulated proteins have been found to be associated with genomic pathways playing an important role in carcinogenesis. In conclusion, the integrated analysis of genomic and proteomic data sets can lead to a better understanding of functional inference at the physiological level and of potential molecular targets in clinical settings through identification of pathways and molecular subnetworks that are implicated in human CRC tissues.

Materials and methods

Tissue samples. CRC tissues and their adjacent normal colonic mucosal tissue counterparts used for 2-nitrobenzenesulfenyl (NBS) labeling were collected from 12 CRC patients who underwent surgical resection at Osaka University Hospital from 2003 to 2004 (clinicopathological features are described in Table I). For DNA microarray, samples of 180 CRC tissues and 40 normal colonic mucosal tissues were obtained from patients who underwent surgical resection at Osaka University Hospital from 2003 to 2006. After surgical resection, tissues were immediately stored at -80˚C prior to analysis. None of the patients received chemotherapy or radiotherapy before surgery. Necrotic tissue was excluded from the study and none of the adenomas contained a cancerous component. All normal tissues were histopathologically confirmed as cancer-free. All patients gave written informed consent; this study was approved by the Institutional Review Board for human tissue use at the Graduate School of Medicine, Osaka University.

Protein separation. Frozen tissue samples were homogenized in 500 ml of lysis buffer A (50 mM Tris-HCl at pH 8.0, 100 mM NaCl, 5 mM EDTA, 1 mM PMSF, 1 mg/ml leupeptin, and 5 mg/ml aprotinin) on ice using a Sample Grinding kit (GE Healthcare, Buckinghamshire, UK). Homogenates were centrifuged at 100,000 x g for 60 min and supernatants were collected as the cytosolic fraction (CF). Pellets were washed twice with lysis buffer A and homogenized in 500 ml of lysis buffer B (2% CHAPS, 9 M urea, 50 mM Tris-HCl at pH 8.0, 100 mM NaCl, 5 mM EDTA, 1 mM PMSF, 1 mg/ml leupeptin, and 5 mg/ml aprotinin); homogenates were centrifuged at 100,000 x g for 60 min. Supernatants were collected as 2% CHAPS-soluble fraction (CSF). These fractionated samples were precipitated using the 2D-Clean-Up kit (Bio-Rad, Hercules, CA) and resuspended in 8 M urea and 5 mM EDTA. After centrifugation at 100,000 x g for 5 min, supernatants were collected and subjected to NBS reagent labeling. Protein concentration was determined using the BCA Protein Assay kit (Pierce, Rockford, IL) using BSA as a standard.

Table I. Characteristics of 12 CRC samples for proteomics and 180 CRC samples for transcriptomics.

	12 CRC samples for proteomics	180 CRC samples for transcriptomics
Gender		
Male	8	111
Female	4	69
Age		
Median (range)	68 (61-78)	67 (30-90)
Location		
Cecum	0	16
A. Colon	3	23
T. Colon	0	10
D. Colon	1	10
S. Colon	6	54
Rectum	2	67
Tumor size (mm)		
Mean ± SD	4.4±1.8	6.7±9.9
Histological type		
Well	1	40
Mod	11	134
Por	0	4
Muc	0	2
Stage		
0/I	2	18
II	5	75
III	5	87
IV	0	0

NBS reagent labeling, peptide fractionation, and MS measurement. NBS reagent labeling was performed according to the manufacturer's protocol (^{13}CNBS stable isotope labeling kit-N; Shimadzu Biotech, Kyoto, Japan). Normal and tumor tissue samples (100 mg each) were labeled with isotopically light and heavy NBS reagents, respectively. NBS-labeled samples were then mixed, reduced, alkylated, and digested by trypsin. NBS-labeled peptides were enriched from tryptic digests and fractionated using Phenyl-Sepharose, as previously described (15). The resulting seven fractions were combined into three fractions and subjected to reversed-phase liquid chromatography (LC-10ADvp mHPLC system; Shimadzu), as previously described (16). Eluates were automatically deposited onto MALDI target plates by the LC spotting system (AccuSpot; Shimadzu). These samples were automatically analyzed by MALDI-TOF MS (AXIMA-CFR Plus; Shimadzu/Kratos, Manchester, UK) (16).

Relative quantification and identification of differentially expressed proteins in HCC. Relative quantification of each NBS-labeled peptide pair was performed using the NBS Analysis Software version 1.0 (Shimadzu), referring to a monoisotopic mass list from MASCOT Distiller version 1.1.2 (Matrix Science), as previously described (16). Peptide pair ratios >1.5 fold or <0.66 fold were set as threshold values. The threshold value for the occurrence was set to 60% of all CRC patient samples in which peptide pairs were detected. In this manner, candidate peptides were selected and further subjected to MS/MS analysis (AXIMAQIT-TOF; Shimadzu/Kratos) (16). Proteins were identified by the MASCOT MS/MS Ion Search algorithm (version 2.0; Matrix Science) using a mass list generated by the MASCOT Distiller. The MASCOT search parameters were: trypsin digestion allowing up to two missed cleavages, fixed modifications of ^{12}CNBS (or ^{13}CNBS) and carbamidomethyl (C), variable modifications of oxidation (M), peptide tolerance 0.3 Da, and MS/MS tolerance of 0.5 Da. Search results with P<0.05 were defined as positive identifications.

Gene expression profiling. Total RNA was purified from 180 CRC tissue samples using TRIzol reagent (Invitrogen, San Diego, CA), as previously described (17). Integrity of the RNA was assessed on the Agilent 2100 Bioanalyzer and RNA 6000 LabChip kits (Yokokawa Analytical Systems, Tokyo, Japan). Only high quality RNA with intact 18s and 28s RNA were used for subsequent analysis. Forty RNA extractions from different normal colonic mucosal tissues were mixed and used as control reference. Extracted RNA samples were amplified with T7 RNA polymerase using the Amino Allyl MessageAmp™ aRNA kit (Ambion, Austin, TX) according to the manufacturer's protocol. The quality of each Amino Allyl-aRNA sample was checked on the Agilent 2100 Bioanalyzer. Control and experimental aRNA samples (5 μg) were labeled with Cy3 and Cy5, respectively; these samples were then mixed, and hybridized on an oligonucleotide microarray covering 30,336 human probes (AceGene Human 30K; DNA Chip Research Inc. and Hitachi Software Engineering Co., Yokohama, Japan). Experimental protocol is available at http://www.dna-chip.co.jp/thesis/AceGeneProtocol.pdf. The microarrays were scanned on the ScanArray 4000 (GSI Lumonics, Billerica, MA). Signal values were calculated using the DNASIS Array Software (Hitachi Software Inc., Tokyo, Japan). Following background subtraction, data with low signal intensities were excluded from additional investigation. In each sample, the Cy5/Cy3 ratio values were log-transformed. Then, global equalization to remove a deviation of the signal intensity between whole Cy3- and Cy5-fluorescence was performed by subtracting the median of all log (Cy5/Cy3) values from each log (Cy5/Cy3) value. Genes with missing values in >20% samples were excluded from further analysis; 24,537 genes out of 30,336 were analyzed.

Statistical integration of proteomics and transcriptomics. To investigate the correlation between mRNA and protein levels for a set of differentially expressed proteins, we used the Spearman rank coefficient and GSEA, a computational method that determines whether an a priori defined set of genes shows statistically significant and concordant differences between two biological states, as previously reported (14). In brief, using the gene expression profiles of the 180 CRC samples, the gene list *L* was ranked by calculating the mean expression level of each gene across all CRC samples. The differentially expressed proteins selected by NBS method were then mapped to their corresponding mRNAs and used as a test group set *S* against the ranked gene list *L* in GSEA analysis. Enrichment score *(ES)* of a test group set *S* characterizes whether the set *S* is randomly distributed across the list or falls mainly at the bottom or top of the list *L*. The null hypothesis that a test group set *S* randomly distributes across the ranked gene list *L* was tested with the Kolmogorov-Smirnov test, and the statistical significance value (nominal P-value) was estimated by 1000 random permutations of the phenotype labels. In this study, the test group set (differentially expressed proteins set) with a P<0.05 of *ES* was considered to be statistically significant correlation between mRNA and protein expression.

Assessment of significantly regulated pathways in CRC. KEGG database (http://www.genome.jp/kegg/pathway.html) and GSEA 2.0.1 (a publicly available desktop application from the Broad Institute (http://www.broad.mit.edu/gsea/software/software_index.html) (14) were employed to assess significantly regulated pathways in the 180 CRC gene expression data sets. In KEGG database, a test set size filter (min=15, max=500) was applied to 200 pathways. The enrichment score *(ES)* of a test set *S* and statistical significance value (nominal P-value) were estimated by permutations of the phenotype labels as described above. Test sets with a high significance (P<0.05) of *ES* were considered as potentially regulated pathways. We normalized the *ES* for each gene set to account for the size of the set, yielding a normalized enrichment score *(NES)*. We then control the proportion of false positives by calculating the false discovery rate (FDR) (18) corresponding to each *NES*. FDR is the estimated probability that a set with a given *NES* represents a false positive finding; it is calculated by comparing the tails of observed and null distributions for *NES*.

Results

Proteomic profiling and identification of differentially expressed proteins in 12 CRC tissues. The NBS method was used for profiling protein expression in a set of tumor and normal tissue samples from 12 CRC patients. After a series of experiments, ~2,600-3,000 peak pairs were observed in each analysis. After relative quantification of each peak pair in each sample, 320 pairs were found to have significant alterations in protein expression and occurred with significant frequency in patients. After these peaks were subjected to MS/MS analysis, 226 MS/MS spectra were obtained, and 156 search results were considered as positive identifications. In total, 128 proteins were confirmed as CRC-associated proteins. Of these, 71 proteins were upregulated and 57 proteins were downregulated in tumor tissue compared to those in normal tissue in CF and CSF analyses, respectively. Using UniPlot Protein database (http://www.uniprot.org/) and matching Entrez Gene ID, we mapped these differentially expressed proteins to their corresponding mRNA counterparts. Finally, we focused on 53 upregulated and 44 downregulated proteins, for which mRNA expression was determined by cDNA microarray analysis (Tables II and III).

Table II. List of 53 upregulated proteins in CRC tissues and transcriptomic expression data.

Upregulated proteins in CRC tissues	Symbol	Entrez Gene ID	Average of 12 CRCs Log_2T/N ratio - proteomics	Average of 180 CRCs Log_2T/N ratio - transcriptomics
α1 acid glycoprotein	ORM1	5004	1.19	-0.29
β-tubulin	TUBB	203068	0.79	0.32
Apurinic endonuclease	APEX1	328	1.00	0.15
Calumenin	CALU	813	1.18	0.32
Chaperonin1	HSPD1	3329	1.10	0.50
Clathrin light polypeptide A	CLTA	1211	1.54	0.13
Complement factor H	CFH	3075	0.88	-0.85
Cysteine rich intestinal protein 1	CRIP1	1396	0.58	-0.30
Cytokeratin 18	KRT18	3875	1.46	0.31
Ezrin	EZR	7430	0.99	-0.38
F-box protein 40	FBXO40	51725	1.30	0.00
Fibrinogen γ	FGG	2266	0.97	-0.07
Galectin 1	LGALS1	3956	0.90	0.83
Glutathione peroxidase 1	GPX1	2876	0.67	0.36
Golgi complex-associated protein 1	ACBD3	64746	1.05	0.12
Heat shock 70 kD protein 9B	HSPA9	3313	1.10	-0.18
Heat shock protein 27	HSPB1	3315	1.14	0.23
Heparan sulfate proteoglycan 2	HSPG2	3339	0.98	0.06
High density lipoprotein-binding protein	HDLBP	3069	0.71	0.22
HLA-C	HLA-C	3107	0.73	-0.35
Hypothetical protein FLJ38663	C12orf65	91574	0.75	-0.38
Inorganic pyrophosphatase	PPA1	5464	1.24	0.67
Mitogen inducible gene 2 protein	FERMT2	10979	0.71	-0.19
Plastin 2	LCP1	3936	1.03	-0.09
Plectin 1	PLEC1	5339	0.84	-0.07
Proteasome subunit p58	PSMD3	5709	0.73	0.21
Pyruvate kinase 3	PKM2	5315	0.93	0.36
RAB22A	RAB22A	57403	0.78	0.11
RACK1	GNB2L1	10399	0.83	0.13
Radixin	RDX	5962	0.84	-0.27
RAN, member RAS oncogene family	RAN	5901	0.83	0.52
Reticulocalbin 1	RCN1	5954	1.32	0.63
Ribosomal protein L13	RPL13	6137	1.48	0.59
Ribosomal protein L27a	RPL27A	6157	0.95	0.20
Ribosomal protein L4	RPL4	6124	0.93	0.16
Ribosomal protein S18	RPS18	6222	1.25	0.26
Ribosomal protein S29	RPS29	6235	0.68	0.10
Ribosome binding protein 1	RRBP1	6238	0.80	0.08
S adenosylhomocysteine hydrolase	AHCY	191	1.22	0.77
S100 calcium binding protein A9	S100A9	6280	1.06	0.68
Solute carrier family 25, member 5	SLC25A5	292	0.84	-0.26
Solute carrier family 3, member 2	SLC3A2	6520	0.85	0.41
Splicing factor 3B, subunit 3	SF3B3	23450	1.24	0.34
Splicing factor, arginine/serine-rich 3 (SRp20)	SFRS3	6428	0.81	-0.13
Transgelin	TAGLN	6876	0.80	0.18
Transgelin 2	TAGLN2	8407	1.00	0.41
Triosephosphate isomerase 1	TPI1	7167	0.99	0.21
Ubiquitin-activating enzyme 1	UBE1	7317	0.90	0.24
U5 snRNP-specific protein, 116 kD	EFTUD2	9343	1.19	0.34
Vimentin	VIM	7431	0.97	-0.09
Vitronectin	VTN	7448	0.70	-0.02
XTP3 transactivatied protein A	XTP3TPA	79077	1.06	0.18
Zyxin	ZYX	7791	1.03	0.63

Spearman rank coefficient between mRNA and protein expression in CRC. Fig. 1 illustrates our experimental design for correlation analysis. We initially investigated the extent to which expression data at the protein and mRNA levels were correlated in CRC using traditional correlation coefficient approach. The identified differentially expressed proteins, the expression ratios for these upregulated and down-regulated proteins, and the expression ratios for each corresponding mRNA transcript are provided in Tables II and III. Fig. 2A and B show the expression ratios of uniquely identified proteins plotted against ratios of the products of the corresponding genes at the mRNA level. A nonparametric correlation analysis of the experimental data using Spearman rank correlation method gave a correlation coefficient of 0.36 for up-regulated proteins data, and a correlation coefficient of 0.28 for down-regulated proteins data. This indicates a poor correlation between mRNA and protein expression ratios.

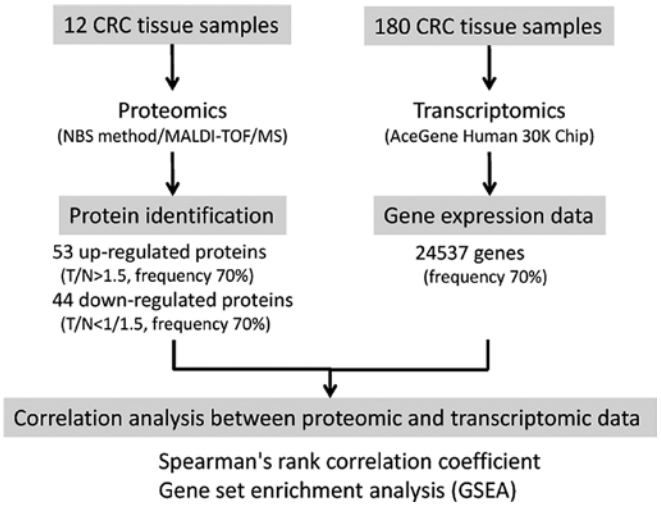

Figure 1. The schematic of our experimental design for correlation analysis. From NBS-based proteomics using 12 CRC tissues, expression data of 53 up-regulated and 44 down-regulated proteins were obtained. From DNA microarray using 180 CRC tissues, 24,536 gene expression profiles were obtained. We investigated the correlation between mRNA and protein levels for a set of differentially expressed proteins, using the Spearman rank coefficient and GSEA.

Evaluation of the correlation between mRNA and protein expression using GSEA. Next, we integrated proteomic data and global transcriptomic data to examine the correlation between mRNA and protein levels using GSEA approach. We focused on the 53 up-regulated and 44 down-regulated proteins shown in Tables II and III. To map these differentially expressed proteins to their corresponding mRNA counterparts, we used the Entrez Gene ID and UniPlot Protein database (http://www.uniprot.org/). Then, 24,537 genes after processing were ranked according to their magnitude of differential gene expression in 180 CRC tissues, obtained by microarray and analyzed using GSEA to examine the distribution of the 53 up-regulated proteins. We found that the 53 up-regulated proteins were significantly enriched in genes exhibiting elevated gene expression levels (P<0.001, ES=0.53), indicating a positive correlation between proteomic and transcriptomic data (Fig. 3A). Similarly, the 44 down-regulated proteins were significantly enriched in genes exhibiting elevated gene expression levels (P<0.001, ES -0.65), indicating a positive correlation between proteomic and transcriptomic data (Fig. 3B).

Identification of characteristic genomic pathways and potential regulatory proteins associated with CRC tissues. After defining the differentially expressed proteins in 180 CRC tissues, the challenge was to objectively interpret potential biological mechanisms. We performed GSEA to identify functional genetic pathways (KEGG database; http://www.genome.jp/kegg/pathway.html) that correlate with the entire ranked gene list from 180 CRC tissues, sorted by their magnitude of differential gene expression. Of the 200 gene sets that contained genes whose products were involved in specific metabolic and signaling pathways, 12 were significantly enriched with nominal P<0.05, as shown in Table IV. Among the molecules within these 12 pathways in KEGG database, we found several up-regulated proteins obtained by NBS analysis. In particular, a relatively large number of up-regulated proteins were related to the principal two pathways; ECM receptor interaction was related to HSPG2 and VTN, and ribosome to RPL13, RPL27A, RPL4, RPS18, and RPS29.

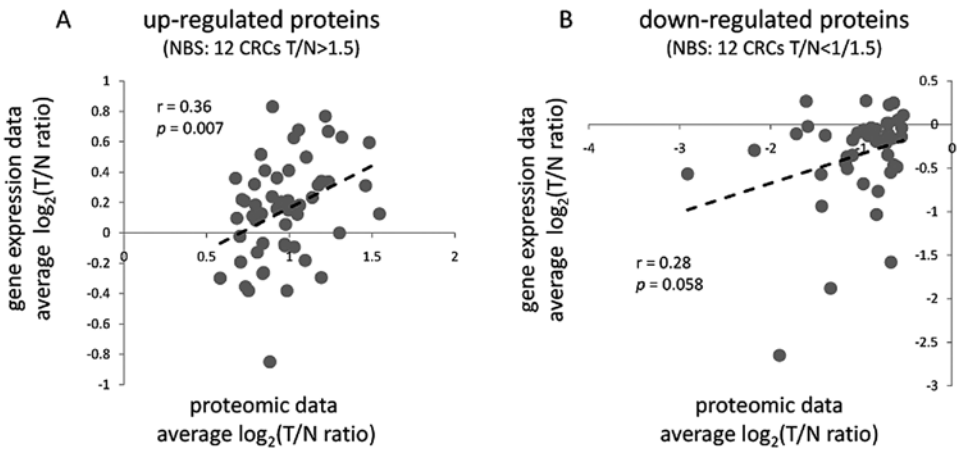

Figure 2. Correlation between mRNA and protein expression levels in CRC, using the Spearman rank coefficient. Expression values are plotted on a log scale. The dotted lines indicate the correlation trend. (A) mRNA-protein correlation for expression values of 53 upregulated proteins. (B) mRNA-protein correlation for expression values of 44 down-regulated proteins.

Table III. List of 44 down-regulated proteins in CRC tissues and trancriptomic expression data.

Down-regulated proteins in CRC tissues	Symbol	Entrez Gene ID	Average of 12 CRCs Log_2T/N ratio - proteomics	Average of 180 CRCs Log_2T/N ratio - transcriptomics
α1 acid glycoprotein	ORM1	5004	1.19	-0.29
ADP ribosylation factor like 10C	ARL8B	55207	-0.97	-0.06
Aldehyde dehydrogenase 2	ALDH2	217	-0.84	-0.05
ATP synthase, H^+ transporting, mitochondrial F0 complex, subunit d	ATP5H	10476	-1.44	-0.57
ATP binding cassette transporter subfamily A member 12	ABCA12	26154	-1.09	-0.18
Calreticulin	CALR	811	-0.54	0.11
Carbonic anhydrase II	CA2	760	-1.90	-2.65
Carbonyl reductase 1	CBR1	873	-0.60	0.05
Cathepsin S	CTSS	1520	-0.81	-0.77
Collagen type 12 α-1	COL12A1	1303	-0.69	0.23
Creatine kinase-B	CKB	1152	-0.67	-1.58
Cysteine rich protein 1	CSRP1	1465	-0.92	-0.13
Dynein light chain 1	DYNLL1	8655	-0.89	-0.04
Endoplasmic-reticulum-lumenal protein 29	ERP29	10961	-0.72	-0.16
Endoplasmic-reticulum-lumenal protein 46	TXNDC5	81567	-1.10	-0.35
Enoyl CoA hydratase 1	ECHS1	1892	-1.18	-0.44
Eukaryotic translation elongation factor 2	EEF2	1938	-0.95	0.27
Eukaryotic translation initiation factor 3 subunit 6	EIF3S6	3646	-1.61	0.27
FHL1 (skeltal muscle LIM-protein)	FHL1	2273	-1.16	-0.37
Gelsolin isoform a	GSN	2934	-0.98	-0.68
Glucosamine-fructose-6-phosphate aminotransferase 1	GFPT1	2673	-0.83	-0.19
GTP-binding protein Rab1	RAB1A	5861	-0.64	-0.46
Haptoglobin	HP	3240	-1.04	-0.10
Heterogeneous nuclear ribonucleoprotein A2 /B1	HNRPA2B1	3181	-0.72	0.02
Hydroxymethylglutaryl-CoA synthase, mitochondrial	HMGCS2	3158	-1.44	-0.94
Isocitrate dehydrogenase 1	IDH1	3417	-2.91	-0.57
JM5 protein	WDR45	11152	-0.64	0.25
Major vault protein	MVP	9961	-0.56	-0.14
Myeloperoxidase	MPO	4353	-0.55	-0.04
Myozenin 3	MYOZ3	91977	-0.73	-0.22
NADH Ubiquinone oxidoreductase subunit B13	NDUFA5	4698	-1.15	-0.50
Normal mucosa of esophagus specific 1	C15orf48	84419	-1.34	-1.88
Olfactomedin 4	OLFM4	10562	-0.61	-0.49
Phosphoenolpyruvate calboxykinase 2	PCK2	5106	-0.68	-0.55
Phosphoglycerate mutase 1	PGAM1	5223	-1.59	-0.02
Proline arginine-rich end leucine-rich repeat protein	PRELP	5549	-1.40	-0.12
Protein kinase C and casein kinase substrate in neurons 2	PACSIN2	11252	-0.66	-0.16
Pyridoxine 5-prime-phosphate oxidase	PNPO	55163	-1.72	-0.11
Ras associated protein Rab5B	RAB5B	5869	-0.70	-0.11
Retinoblastoma binding protein 4, 7	RBBP4/RBBP7	5928/5931	-0.70	0.02
Succinate dehydrogenase complex, subunit A, flavoprotein	SDHA	6389	-0.71	-0.35
Transferrin	TF	7018	-0.85	-0.11
UDP-glucose dehydrogenase	UGDH	7358	-0.83	-1.03
Valosin containing protein	VCP	7415	-0.95	-0.09
Villin 1	VIL1	7429	-2.18	-0.30

Discussion

The recent availability of platform technologies for high throughput transcriptomics and proteomics has led to integrated approaches to cancer research. Integrated analysis of global scale transcriptomics and proteomics data can provide important insights into the biological mechanisms underlying complex physiological processes (19). However, it is difficult to accurately evaluate their correlation using the conventional correlation coefficient analysis, which deals with a fold change

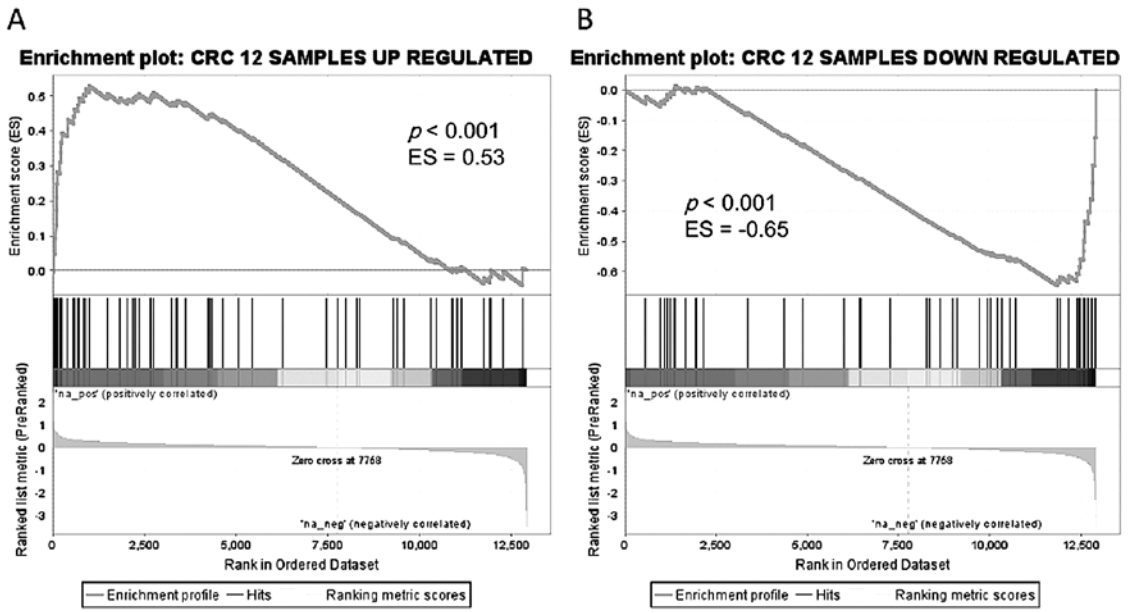

Figure 3. Enrichment plots for (A) 53 up-regulated proteins and (B) 44 down-regulated proteins in 12 CRC tissues. Top, the running enrichment score for the protein set, as the analysis sweeps through the entire ranked list organized on the basis of their magnitude of differential gene expression in 180 CRC tissues. The score at the peak of the plot is the enrichment score (ES) for the protein set. Middle, members of the protein set appear in a ranked list of genes. Bottom, the value of the ranking metric along the list of ranked genes.

Table IV. Significantly enriched genomic pathways of the KEGG database.

KEGG pathways	Size	NES	NOM P-val	FDR q-val	NBS up-regulated proteins
ECM receptor interaction	73	2.44	<0.001	<0.001	HSPG2, VTN
Ribosome	59	2.40	<0.001	<0.001	RPL13, RPL27A, RPL4, RPS18, RPS29
Cell communication	99	2.13	<0.001	<0.001	
Focal adhesion	167	1.97	<0.001	0.003	VTN, ZYX
Toll-like receptor signaling pathway	78	1.78	<0.001	0.028	
RNA polymerase	18	1.76	0.006	0.029	
Cell cycle	100	1.67	0.004	0.068	
Proteasome	22	1.65	0.023	0.071	PSMD3
WNT signaling pathway	117	1.56	<0.001	0.120	
Aminoacyl-tRNA biosynthesis	36	1.47	0.040	0.225	
Folate biosynthesis	36	1.42	0.050	0.245	
Cytokine cytokine receptor interaction	187	1.42	0.014	0.232	

GSEA was performed using gene sets from KEGG database. List of the top 12 gene sets enriched in the ranked gene list of 180 CRCs with nominal P-value <0.05. The gene list is sorted by FDR q-val in ascending order. ES; enrichment score, NES; normalized enrichment score, FDR; false discovery rate.

level in individual gene and protein expression. Most recent studies have either failed to find a correlation between protein and mRNA abundance (6) or have observed only a weak correlation (9,12). The reason is that the correlation between protein abundance and mRNA expression level depends on various biological processes and technical factors. With regard to biological processes, transcription and translation do not have a linear and simple relationship (20). Regulatory proteins and sRNAs also act as translational modulators (21). The half-live of an individual protein and the protein turnover are undoubtedly influencing the correlation between mRNA and protein expression to a considerable degree (22).

In this study, we assumed that development of a sophisticated statistical approach was essential to overcome these limitations. To assess the correlation between mRNA and protein expression, we tried the novel approach of GSEA that deals with entire genes represented by an array as ranked gene list ordered by phenotypic correlation (14,23). As expected, the conventional Spearman rank coefficient showed only a weak correlation. In GSEA, when differentially expressed proteins were treated as

a collective set, significant concordance was observed with primary tumor gene expression data.

In the present study, NBS method also plays a role in improving the precision of correlation analysis. This method is based on stable isotope labeling of tryptophan residues by NBS reagents. As previously described (15,24), this novel method has two major advantages: it reduces the number of peptides by selecting NBS-labeled tryptophan-containing peptides from bulk tryptic digests and a special matrix is used for MALDI-TOF MS measurements that can detect NBS-labeled peptides with high sensitivity. Therefore, this method could improve proteome mining by increasing the dynamic range of detection, and it shows potential for quantitative proteome analysis (25,26).

As a following step, GSEA enabled us to identify 12 potentially regulated genetic pathways in CRC. Cell-ECM (extracellular matrix) interaction is an essential mechanism in several biological processes, such as cell proliferation, migration, differentiation, apoptosis, as well as carcinogenesis (27,28). CRC cells invade the stroma as coherent cell nests instead of single cells (29,30). Of the molecules within ECM interaction pathway, HSPG2 and VTN were revealed as potential key modulators by NBS-based proteomics. HSPG2, the large prominent heparan sulfate proteoglycan of extracellular matrices, is known as a component that may participate in ECM interaction (31). Within the matrix, vitronectin can support cellular adhesion through interactions with integrins (32). In addition, vitronectin is a major component of the stroma of primary hepatocellular carcinoma and metastatic hepatic tumors including colorectal hepatic metastases (32,33).

The ribosome pathway is essential for protein synthesis. The increased overall ribosome biogenesis is a well-known common feature of active proliferation, and the proliferation rate of tumor cells is higher than that of normal ones (34). Ribosomal proteins (RPs) showed different expression patterns: not all RPs increased in the same tumor or tissue, and the same RP was expressed differentially in different tumors or different stages of diseases (35). In our study, of the molecules within Ribosomal proteins, RPL13, RPL27A, RPL4, RPS18, and RPS29 were up-regulated at the protein level. Previous studies on CRC revealed extraribosomal functions for these proteins including self-translation regulation, development regulation, and tumor suppressor gene regulation (36,37).

The assumption of GSEA is that functional gene sets with significantly high expression coherence suggest putative functionality. It must be noted that annotated functions of gene sets with higher expression coherence do not always directly correspond with the actual biological functions (38). Nonetheless, several physiological cellular responses require simultaneous participation of gene products, and genes with central roles are likely to have similar regulatory control and expression patterns (39,40). Comparative analysis also showed that coexpression patterns of many functionally related genes are conserved across diverse species (41). Thus, most gene sets with significantly high expression coherence, if not all, might represent key molecular functions of the corresponding expression profiles.

In conclusion, we performed an integrated analysis of mRNA and protein expression data in CRC. Overall, significant correlation was observed between changes in mRNA and protein levels that were consistent with the expectation that a substantial proportion of changes in protein would be a consequence of changes in mRNA levels rather than post-transcriptional effects. Our identified regulatory signatures of mRNA and protein levels might be able to enhance the understanding of carcinogenesis and cancer proliferation and lead to the elucidation of novel molecular targets in the clinical field.

Acknowledgments

The authors thank Kenichi Matsubara, Ph.D. (President and Executive Director of The DNA Chip Research Inc., Yokohama Japan) for contract services on AceGene Human 30K.

Refernces

1. Acharya CR, Hsu DS, Anders CK, *et al*: Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. JAMA 299: 1574-1587, 2008.
2. Wulfkuhle JD, Liotta LA and Petricoin EF: Proteomic applications for the early detection of cancer. Nat Rev Cancer 3: 267-275, 2003.
3. Ransohoff DF: Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 4: 309-314, 2004.
4. Yamasaki M, Takemasa I, Komori T, *et al*: The gene expression profile represents the molecular nature of liver metastasis in colorectal cancer. Int J Oncol 30: 129-138, 2007.
5. Watanabe M, Takemasa I, Kawaguchi N, *et al*: An application of the 2-nitrobenzenesulfenyl method to proteomic profiling of human colorectal carcinoma: A novel approach for biomarker discovery. Proteomics Clin Appl 2: 925-935, 2008.
6. Gygi SP, Rochon Y, Franza BR and Aebersold R: Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720-1730, 1999.
7. Chen G, Gharib TG, Huang CC, *et al*: Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 1: 304-313, 2002.
8. Pradet-Balade B, Boulme F, Beug H, Mullner EW and Garcia-Sanz JA: Translation control: bridging the gap between genomics and proteomics? Trends Biochem Sci 26: 225-229, 2001.
9. Greenbaum D, Colangelo C, Williams K and Gerstein M: Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 4: 117, 2003.
10. Beyer A, Hollunder J, Nasheuer HP and Wilhelm T: Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. Mol Cell Proteomics 3: 1083-1092, 2004.
11. Ideker T, Thorsson V, Ranish JA, *et al*: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929-934, 2001.
12. Washburn MP, Koller A, Oshiro G, *et al*: Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 100: 3107-3112, 2003.
13. Futcher B, Latter GI, Monardo P, McLaughlin CS and Garrels JI: A sampling of the yeast proteome. Mol Cell Biol 19: 7357-7368, 1999.
14. Subramanian A, Tamayo P, Mootha VK, *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545-15550, 2005.
15. Matsuo E, Toda C, Watanabe M, *et al*: Improved 2-nitrobenzenesulfenyl method: optimization of the protocol and improved enrichment for labeled peptides. Rapid Commun Mass Spectrom 20: 31-38, 2006.
16. Iida T, Kuyama H, Watanabe M, *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 17: 333-341, 2006.
17. Kittaka N, Takemasa I, Takeda Y, *et al*: Molecular mapping of human hepatocellular carcinoma provides deeper biological insight from genomic data. Eur J Cancer 44: 885-897, 2008.
18. Reiner A, Yekutieli D and Benjamini Y: Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375, 2003.

19. Alter O and Golub GH: Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. Proc Natl Acad Sci USA 101: 16577-16582, 2004.
20. Nahvi A, Barrick JE and Breaker RR: Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. Nucleic Acids Res 32: 143-150, 2004.
21. Golding I, Paulsson J, Zawilski SM and Cox EC: Real-time kinetics of gene activity in individual bacteria. Cell 123: 1025-1036, 2005.
22. Doherty MK, Hammond DE, Clague MJ, Gaskell SJ and Beynon RJ: Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. J Proteome Res 8: 104-112, 2009.
23. Mootha VK, Lindgren CM, Eriksson KF, *et al*: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267-273, 2003.
24. Matsuo E, Toda C, Watanabe M, *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 6: 2042-2049, 2006.
25. Ou K, Kesuma D, Ganesan 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 5: 2194-2206, 2006.
26. Ueda K, Katagiri T, Shimada T, *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 6: 3475-3483, 2007.
27. Fischer H, Stenling R, Rubio C and Lindblom A: Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. Carcinogenesis 22: 875-878, 2001.
28. Meyer S, Hafner C, Guba M, *et al*: Ephrin-B2 overexpression enhances integrin-mediated ECM-attachment and migration of B16 melanoma cells. Int J Oncol 27: 1197-1206, 2005.
29. Shimao Y, Nabeshima K, Inoue T and Koono M: Complex formation of IQGAP1 with E-cadherin/catenin during cohort migration of carcinoma cells. Its possible association with localized release from cell-cell adhesion. Virchows Arch 441: 124-132, 2002.
30. Nabeshima K, Shimao Y, Inoue T and Koono M: Immunohistochemical analysis of IQGAP1 expression in human colorectal carcinomas: its overexpression in carcinomas and association with invasion fronts. Cancer Lett 176: 101-109, 2002.
31. Hummel S, Osanger A, Bajari TM, *et al*: Extracellular matrices of the avian ovarian follicle. Molecular characterization of chicken perlecan. J Biol Chem 279: 23486-23494, 2004.
32. Edwards S, Lalor PF, Tuncer C and Adams DH: Vitronectin in human hepatic tumours contributes to the recruitment of lymphocytes in an alpha v beta3-independent manner. Br J Cancer 95: 1545-1554, 2006.
33. Jaskiewicz K, Chasen MR and Robson SC: Differential expression of extracellular matrix proteins and integrins in hepatocellular carcinoma and chronic liver disease. Anticancer Res 13: 2229-2237, 1993.
34. Pogue-Geile K, Geiser JR, Shu M, *et al*: Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. Mol Cell Biol 11: 3842-3849, 1991.
35. Seshadri T, Uzman JA, Oshima J and Campisi J: Identification of a transcript that is down-regulated in senescent human fibroblasts. Cloning, sequence analysis, and regulation of the human L7 ribosomal protein gene. J Biol Chem 268: 18474-18480, 1993.
36. Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M and Quattrone A: Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. Carcinogenesis 27: 1323-1333, 2006.
37. Kimura K, Wada A, Ueta M, *et al*: Comparative proteomic analysis of the ribosomes in 5-fluorouracil resistance of a human colon cancer cell line using the radical-free and highly reducing method of two-dimensional polyacrylamide gel electrophoresis. Int J Oncol 37: 1271-1278, 2010.
38. Pavlidis P, Lewis DP and Noble WS: Exploring gene expression data with class scores. Pac Symp Biocomput 2002: 474-485, 2002.
39. Segal E, Wang H and Koller D: Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19 (Suppl 1): i264-i271, 2003.
40. Graeber TG and Eisenberg D: Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. Nat Genet 29: 295-300, 2001.
41. Stuart JM, Segal E, Koller D and Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249-255, 2003.