SPANDIDOS PUBLICATIONS

Genome-wide molecular characterization of mucinous colorectal adenocarcinoma using cDNA microarray analysis

HAN SANG KIM[1,2], SEUNG HUI KANG[1], CHAN HEE PARK[1], WOO ICK YANG[3], HEI CHEUL JEUNG[1,2], HYUN CHEOL CHUNG[1,2], JAE KYUNG ROH[2], JOONG BAE AHN[2], NAM KYU KIM[4], BYUNG SOH MIN[4] and SUN YOUNG RHA[1,2]

[1]Cancer Metastasis Research Center, Yonsei Cancer Center, Departments of [2]Internal Medicine, [3]Pathology and [4]Surgery, Yonsei University Health System, Seoul, Republic of Korea

Received September 9, 2010; Accepted October 25, 2010

DOI: 10.3892/or.2010.1126

Abstract. Mucinous colorectal carcinoma exhibits distinct clinicopathological features compared to non-mucinous colorectal carcinoma. Previous studies have discovered several molecular genetic features in mucinous colorectal carcinomas, but have limitations as they are confined to a small number of molecules. To understand the mucinous colorectal carcinoma system, this study was designed to identify genes that are differentially expressed in mucinous colorectal carcinoma compared to non-mucinous colorectal carcinoma using cDNA microarrays. cDNA microarray experiments were performed using human cDNA 17k chips with 25 mucinous and 27 non-mucinous cancer tissues. Differentially expressed genes (DEGs) were determined by Welch's t-test and more accurate classifiers were selected from the DEGs using the prediction analysis for microarrays (PAM) software package. Array results were validated using quantitative real-time RT-PCR. The identified gene set was functionally investigated through *in silico* analysis. Sixty-two DEGs were identified and the 50 highest ranking genes could be used to accurately classify mucinous and non-mucinous colorectal carcinomas. The identified gene set included up-regulated *TFF1* (4-fold), *AGR2* (3.3-fold), *FSCN1* (2.2-fold), *CD44* (1.5-fold) and down-regulated *SLC26A3* (0.2-fold) in MC. *TFF1*, *AGR2* and *SLC26A3* were validated by quantitative real-time RT-PCR. The functions of these DEGs were related to tumorigenesis (14 genes), cell cycle progression (6 genes), invasion (2 genes), anti-apoptosis (7 genes), cell adhesion and proliferation (5 genes) and carbohydrate metabolism (3 genes). We suggest that MC has distinct molecular characteristics from NMC and therefore, that the expression signatures of DEGs may improve the understanding of molecular pathogenesis and clinical behaviors in MC.

Correspondence to: Dr Sun Young Rha, Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul 120-752, Republic of Korea
E-mail: rha7655@yuhs.ac

Key words: mucinous carcinoma, colorectal cancer, classification analysis, cDNA microarray, *in silico* analysis

Introduction

Mucinous colorectal carcinoma (MC) is a subtype of colorectal adenocarcinoma that is characterized by an extracellular mucin content of more than 50% of the tumor volume. MC has distinctive clinical and molecular features compared to non-mucinous colorectal carcinoma (NMC). Clinically, mucinous colorectal carcinoma tends to occur more frequently in patients who are less than 50 years old (1,2), be located in the right colon (1,2), present at an advanced stage, invade the adjacent viscera, have more extensive lymph node involvement (3), and have a worse overall 5-year survival rate than NMC (4,5). Several studies have shown that genetic alterations and factors related to microsatellite instability (MSI) (6-9), CpG island methylator phenotype (CIMP) (9), *BRAF* mutation (8-10), MUC2 and MUC5AC (11,12) are related to MC.

Microarray technology has been adapted to profile thousands of genes simultaneously, and has demonstrated the potential use of expression profiles for the genome-wide molecular classification of cancer (13,14). We used a cDNA microarray technique to identify molecular features that discriminate MC and NMC. MC has distinct molecular features compared with NMC, and we suggest that these features may underlie the different cancer characteristics of colorectal cancer subtypes.

Materials and methods

Colorectal adenocarcinoma tissues. Fresh frozen tissues were obtained from colorectal cancer (CRC) patients who underwent curative surgical resection at the Yonsei Cancer Center, Severance Hospital in Seoul, Korea from 2003 to 2006. Pathologists at Severance Hospital strictly evaluated the histologies and mucin volumes of all tumors. Tumors were considered to be MC when mucin covered ≥50% of the microscopically observed areas. Tumors with mucin in 10% of the observed fields were classified as NMC. The tumors with 11-49% mucin content were classified as intermediate mucinous carcinoma (IMC). Clinical information was collected

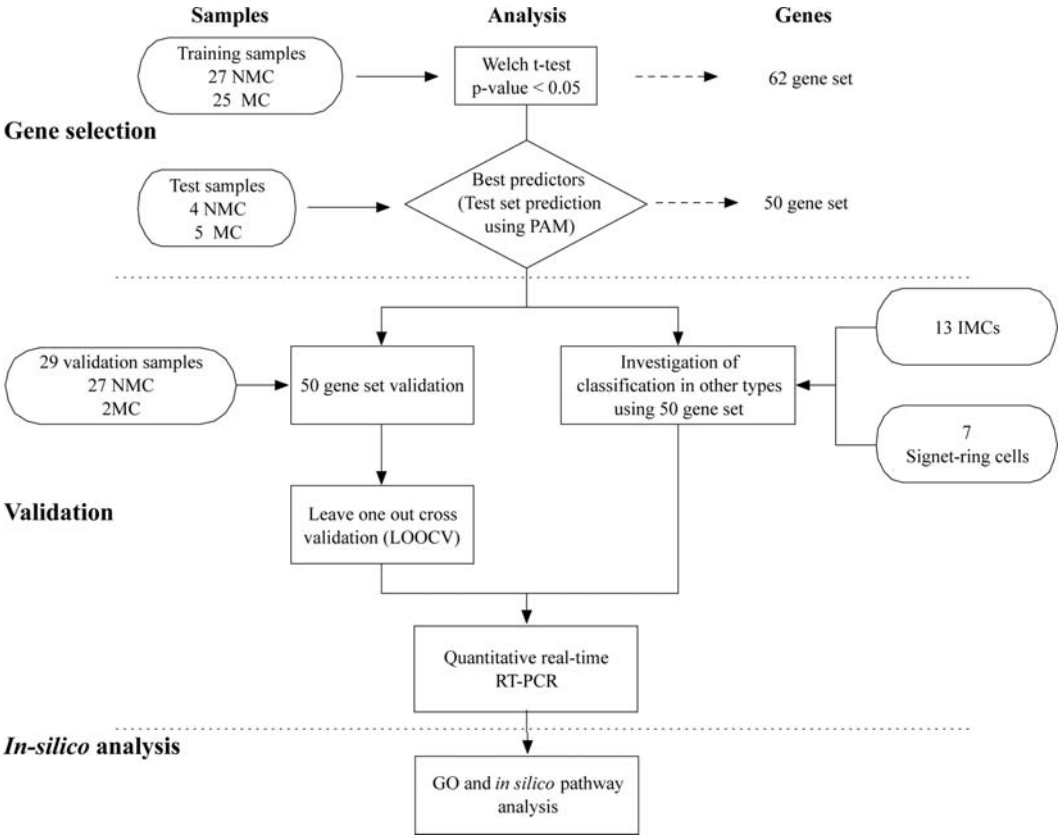

Figure 1. The three stages of the analysis scheme: gene selection, validation and functional gene annotation.

from institutional medical records for all patients. The study protocol was approved by the Severance Hospital Institutional Review Board, and informed consent was obtained from patients for use of surgical specimens and clinicopathologic data for research purposes.

RNA preparation, amplification and cDNA microarray experiments. Total RNA was extracted from tissues using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and amplified using the T7 linear amplification method according to the manufacturer's instructions. Yonsei reference RNA [Cancer Metastasis Research Center (CMRC), Seoul, Korea] was prepared by pooling equivalent amounts of the total RNA from 11 human cancer cell lines of various tumor types (15). The cDNA microarray analysis was performed using a human cDNA 17k chip (CMRC-GT, Seoul, Korea) containing 17104 known genes and ESTs in a reference design following institutional protocols (15).

Genomic DNA extraction and mutation analysis. Genomic DNA was extracted from tissue samples. The sequences of the *BRAF* and *KRAS* mutations were analyzed using a PyroMarkTMID sequencing machine and PyroMarkTMID 1.0 software (Biotage AB and Biosystems, Uppsala, Sweden). The analysis protocol and options were performed as per the manufacturer's recommendations. The *BRAF* gene mutation at position 600 (BRAFV600E) was assessed as described by a previous report (16). Mutations in codon 12 and 13 of the *KRAS* gene were also determined by the Pyrosequencing technology with a slight modification of the previously reported method (8).

Processes of sample selection and analysis. The processes of sample selection and analysis are summarized in Fig. 1. Samples were divided into three sets, training, test and validation, for the selection of classifiers to discriminate MCs and NMCs. Twenty-five MCs were randomly selected from the CRC samples. Twenty-seven matched NMC samples having similar clinicopathologic factors with MCs were then selected and these 52 samples were used as a training set. Welch's t-test was performed to find DEGs discriminating MCs from NMCs in the training set. To identify better predictive genes, we used Prediction Analysis for Microarray (PAM) software (http://www-stat.stanford.edu/~tibs/PAM/) to conduct another test set consisting of 5 MCs and 4 NMCs. Finally, gene annotation of selected probes as classifiers was performed using the SOURCE database (http://source.stanford.edu). The predictive power of the selected gene set was validated using an independent validation set consisting of 27 NMCs and 2 MCs. Additionally, the robust prediction of classifiers for larger data sets, which consisted of the training, test, and validation sets, was confirmed using leave-one-out cross validation (LOOCV). For additional validation, IMC and signet-ring cell colorectal carcinoma samples were investigated by PAM analysis to determine the possibility of classification according to mucin volume. The validation of microarray results was conducted using real-time RT-PCR.

Statistical analysis and class prediction analysis. To identify DEGs, Welch's t-test was applied using GeneSpring GX 7.3.1. The options of the test were set to: FDR<0.05, parametric test, not equal variance and multiple testing correction of Benjamini and Hochberg False Discovery Rate. A χ^2 test was

ample information.

		Training set[b]	
	Total patients[a]	Non-mucinous adenocarcinoma	Mucinous adenocarcinoma
No. of patients	110	27	25
Gender (male/female)	60/50	14/13	16/9
Median age, years (range)	59.5 (30-94)	63 (30-94)	61 (33-78)
Mucin volume			
≤10% (NMC)	58	27	0
≤50% (IMC)	13	0	0
>50% (MC)	39	0	25
Tumor site			
Colon	72	15	16
Rectum	38	12	9
Primary location			
Cecum	16	2	3
Ascending	21	3	4
Transverse	7	3	3
Descending	2	0	0
Sigmoid	26	7	6
Rectum	38	12	9
Tumor depth			
T2	2	0	0
T3	91	25	20
T4	17	2	5
TNM stage (MAC stage)			
II (B)	37	10	7
III (C)	59	13	14
IV (D)	14	4	4

NMC, non-mucinous adenocarcinoma; IMC, intermediate mucinous adenocarcinoma; MC, mucinous adenocarcinoma. [a]Total samples include NMCs, MCs, 13 IMCs and 7 signet-ring cells. [b]In the training set, the NMC samples were matched with MCs according to clinicopathologic factors and none of these factors were significantly different between MCs and NMCs.

utilized for the analysis of categorical data. PAM software v1.30.0 for R (pamr) and all processes for the predictions using PAM were conducted using scripts of pamr on R Cocoa GUI 1.16 (http://www.r-project.org/index.html). Prediction analyses using the Support Vector Machine (SVM) were performed with GeneSpring GX 7.3.1. Options for the SVM were as follows: i) gene selection method: all genes from selected list, ii) kernel function: polynorminal dot product (order 1 and 3) diagonal scaling factor, 10. LOOCV was executed using the GeneSpring GX 7.3.1 internal script and all options were set as those of SVM. The principal component analysis (PCA) was conducted using GeneSpring GX 7.3.1. The PCA on conditions was conducted and the mean centering and scaling method was used. Hierarchical clustering was performed using GeneSpring with complete linkage and Pearson correlation distance.

Quantitative real-time RT-PCR. TFF1, *AGR2*, *SLC26A3* and *MUC2* were selected for validation of the microarray data.

Table II. Prediction results of the test and validation sets.

	Test set (n=9) (%)		Validation set (n=29) (%)	
Gene set[a]	PAM[b]	SVM[c]	PAM	SVM
62 genes	8/9 (88.9)	8/9 (88.9)	23/29 (79.3)	22/29 (75.9)
50 genes	9/9 (100)	9/9 (100)	25/29 (86.2)	23/29 (79.3)

[a]Sixty-two genes were initially selected through Welch's t-test. [b]PAM, prediction analysis for microarray. [c]SVM, support vector machine. Fifty genes were then determined from 62 genes via PAM analysis.

Quantitative real-time RT-PCR (qRT-PCR) was performed on 26 randomly selected samples from the 52 sample training set. Each reaction was run in duplicate using a Stratagene MX3005P Real-Time PCR System (Stratagene, La Jolla, CA,

Table III. Genes that are differentially expressed between MCs and NMCs.

Order[a]	GenBank accession no.	Description	Symbol	Fold change (MC/NMC)[b]	FDR[c]
1	AW082097	Peptidase inhibitor 3, skin-derived (SKALP)	PI3	1.73	0.0052
2	AW073291	Anterior gradient homolog 2 (*Xenopus laevis*)	AGR2	1.71	0.0055
3	AW009769	Trefoil factor 1	TFF1	2.01	0.0180
4	AA490263	NIMA (never in mitosis gene a)-related kinase 3	NEK3	-0.85	0.0028
5	H10045	Vav 3 oncogene	VAV3	-1.27	0.0053
6	AA933744	LINE-1 type transposase domain containing 1	L1TD1	1.56	0.0393
7	AA598652	Transcribed locus		-1.06	0.0062
8	AA412284	Poliovirus receptor	PVR	-0.81	0.0033
9	R09561	CD55 molecule, decay accelerating factor for complement (Cromer blood group)	CD55	0.82	0.0468
10	AW072118	Fascin homolog 1, actin-bundling protein (*Strongylocentrotus purpuratus*)	FSCN1	1.15	0.0036
11	AA931716	Tripartite motif-containing 7	TRIM7	0.91	0.0231
12	AI298104	Amylo-1, 6-glucosidase, 4-α-glucanotransferase (glycogen debranching enzyme, glycogen storage disease type III)	AGL	-0.61	0.0004
13	AI369785	Neural proliferation, differentiation and control, 1	NPDC1	0.76	0.0220
14	AA453310	α-methylacyl-CoA racemase	AMACR	-0.66	0.0294
15	AA279145	Calcium binding protein 39-like	CAB39L	-0.98	0.0136
16	AI339538	Solute carrier family 26, member 3	SLC26A3	-2.47	0.0380
17	AI350851	EST		0.72	0.0028
18	W47576	N-acylsphingosine amidohydrolase (acid ceramidase)-like	ASAHL	-0.66	0.0231
19	AA488868	Acyltransferase like 2	AYTL2	0.74	0.0052
20	AA995282	Four and a half LIM domains 2	FHL2	0.53	0.0421
21	AI700308	Protein phosphatase 1, regulatory (inhibitor) subunit 3D	PPP1R3D	-0.62	0.0055
22	AA127096	Fusion [involved in t(12;16) in malignant liposarcoma]	FUS	-0.49	0.0342
23	AA155640	Transcobalamin I (vitamin B12 binding protein, R binder family)	TCN1	1.46	0.0164
24	AI990501	Stathmin-like 2	STMN2	-0.54	0.0342
25	AI349090	Full-length cDNA clone CS0DI022YE21 of Placenta Cot 25-normalized of *Homo sapiens* (human)		0.64	0.0342
26	N74236	Membrane protein, palmitoylated 1, 55 kDa	MPP1	-0.86	0.0342
27	H87106	EST		-0.80	0.0442
28	AA490497	Ubiquitin-like 3	UBL3	-0.57	0.0056
29	H04789	Glycogenin 2	GYG2	-0.67	0.0393
30	AA040387	X-prolyl aminopeptidase (aminopeptidase P) 2, membrane-bound	XPNPEP2	-0.46	0.0056
31	AI203404	Transcribed locus		0.73	0.0342
32	AI359768	Ras homolog gene family, member U	RHOU	-0.64	0.0342
33	AA042990	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	SEMA3C	-0.82	0.0250
34	AA459012	LMBR1 domain containing 1	LMBRD1	-0.56	0.0140
35	AA282208	EST		-0.56	0.0220
36	AA976544	Melanophilin	MLPH	0.73	0.0342
37	AI688155	Selenophosphate synthetase 2	SEPHS2	-0.58	0.0263
38	AA278698	Haloacid dehalogenase-like hydrolase domain containing 1A	HDHD1A	-0.61	0.0245
39	AA181085	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3	KDELR3	0.58	0.0417
40	AI269774	Phytanoyl-CoA 2-hydroxylase	PHYH	-0.51	0.0380
41	AA481464	Peptidylprolyl isomerase B (cyclophilin B)	PPIB	0.62	0.0058
42	AI051281	Chromosome 13 open reading frame 23	C13orf23	-0.53	0.0417

 Continued..

Order[a]	GenBank accession no.	Description	Symbol	Fold change (MC/NMC)[b]	FDR[c]
43	AA872095	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, ß polypeptide	YWHAB	-0.49	0.0058
44	AA488433	Slowmo homolog 2 (*Drosophila*)	SLMO2	-0.71	0.0157
45	AI989542	CCAAT/enhancer binding protein (C/EBP), α	CEBPA	-0.42	0.0056
46	AI360206	Chromosome 20 open reading frame 112	C20orf112	-0.55	0.0294
47	AA283090	CD44 molecule (Indian blood group)	CD44	0.61	0.0342
48	H73234	CDC42 effector protein (Rho GTPase binding) 1	CDC42EP1	0.44	0.0417
49	AA669068	Staufen, RNA binding protein, homolog 1 (*Drosophila*)	STAU1	-0.48	0.0192
50	AI393075	Cytochrome P450, family 4, subfamily F, polypeptide 3	CYP4F3	-0.50	0.0076
51	AI654481	Chloride channel 2	CLCN2	-0.36	0.0421
52	AA282253	Hect domain and RLD 3	HERC3	-0.47	0.0136
53	H14359	E74-like factor 4 (ets domain transcription factor)	ELF4	-0.46	0.0342
54	AA918102	Transcribed locus, moderately similar to XP_001072659.1 similar to Dolichol-phosphate mannosyltransferase (Dolichol-phosphate mannose synthase) (Dolichyl-phosphate ß-D-mannosyltransferase) (Mannose-P-dolichol synthase) (MPD synthase) (DPM synthase) (*Rattus norvegicus*)		-0.46	0.0423
55	AA465396	Solute carrier family 25, member 37	SLC25A37	0.51	0.0270
56	AA017125	Sortilin 1	SORT1	0.34	0.0421
57	AI632018	Tubulointerstitial nephritis antigen	TINAG	-0.35	0.0500
58	AI769855	Defensin, ß 1	DEFB1	-0.54	0.0294
59	AA458870	Cell division cycle 37 homolog (*S. cerevisiae*)	CDC37	0.30	0.0421
60	AA446884	Tripartite motif-containing 13	TRIM13	-0.32	0.0500
61	AI990104	Microtubule-associated protein, RP/EB family, member 1	MAPRE1	-0.37	0.0482
62	AA633997	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, θ polypeptide	YWHAQ	-0.34	0.0342

Sixty-two genes were selected by Welch's t-test under a P-value threshold of 0.05. The final 50 genes selected with PAM analysis were exactly matched with the top 50 genes. [a]The order of the 62 genes was concluded from the class prediction analysis using PAM, which was calculated based on the standardized centroids of genes for each class. [b]The fold change was calculated as a base 2 logarithm of (average of normalized R/G ratio in MC samples)/(average of normalized R/G ratio in NMC samples). [c]The false discovery rate (FDR) was calculated from Welch's t-test between MC and NMC samples with the Benjamini and Hochberg multiple testing correction.

USA). In brief, 4 μg of amplified RNA from each sample was reverse-transcribed using SuperScript II reverse transcriptase and random primers (Invitrogen). An aliquot of single stranded cDNA from each reverse-transcribed sample (1.4 μl) was PCR amplified using QuantiTect SYBR Green PCR (Qiagen, Valencia, CA, USA). Expression values for each gene were determined using a standard curve constructed from human genomic DNA (Promega, Madison, WI, USA). The housekeeping gene ACTB was selected for normalization and to construct a standard curve. Non-template-control wells without cDNA were included as negative controls. The primer sets for PCR amplification were designed as follows: *TFF1*-F: 5'-TTGTGGTTTTCCTGGTGTCA-3', *TFF1*-R: 5'-CCGAGCTCTGGGACTAATCA-3', *AGR2*-F: 5'-TCCCTTCCTTGAGCATTTTG-3', *AGR2*-R: 5'-GGCCTTGAGACTTGAAACCA-3', *SLC26A3*-F: 5'-TGGCGCCACTATACTGCTAA-3', *SLC26A3*-R: 5'-TTCAAACTTTGGAACAAGATGG-3', *MUC2*-F: 5'-TGGAAAGCAAGGACTGAACA-3', *MUC2*-R: 5'-TACACCCACATCGAGAGCTG-3'. Student's t-test was used to assess the statistical differences in gene expression levels measured by qRT-PCR between the NMC and MC groups.

Prediction of biological functions. Investigation of biological roles was accomplished through the use of ingenuity pathways analysis (IPA, Ingenuity® Systems, www.ingenuity.com). To construct a molecular network related to the determined functions, 11622 probes in the chip were used. The fold ratio (MCs/NMCs) for each probe was calculated, and then these values were parsed from the GeneSpring GX 7.3 to the IPA via the GeneSpring GX-IPA connector scripts.

Results

The selection of the training set and baseline characteristics. To eliminate the effects of factors other than gene expression,

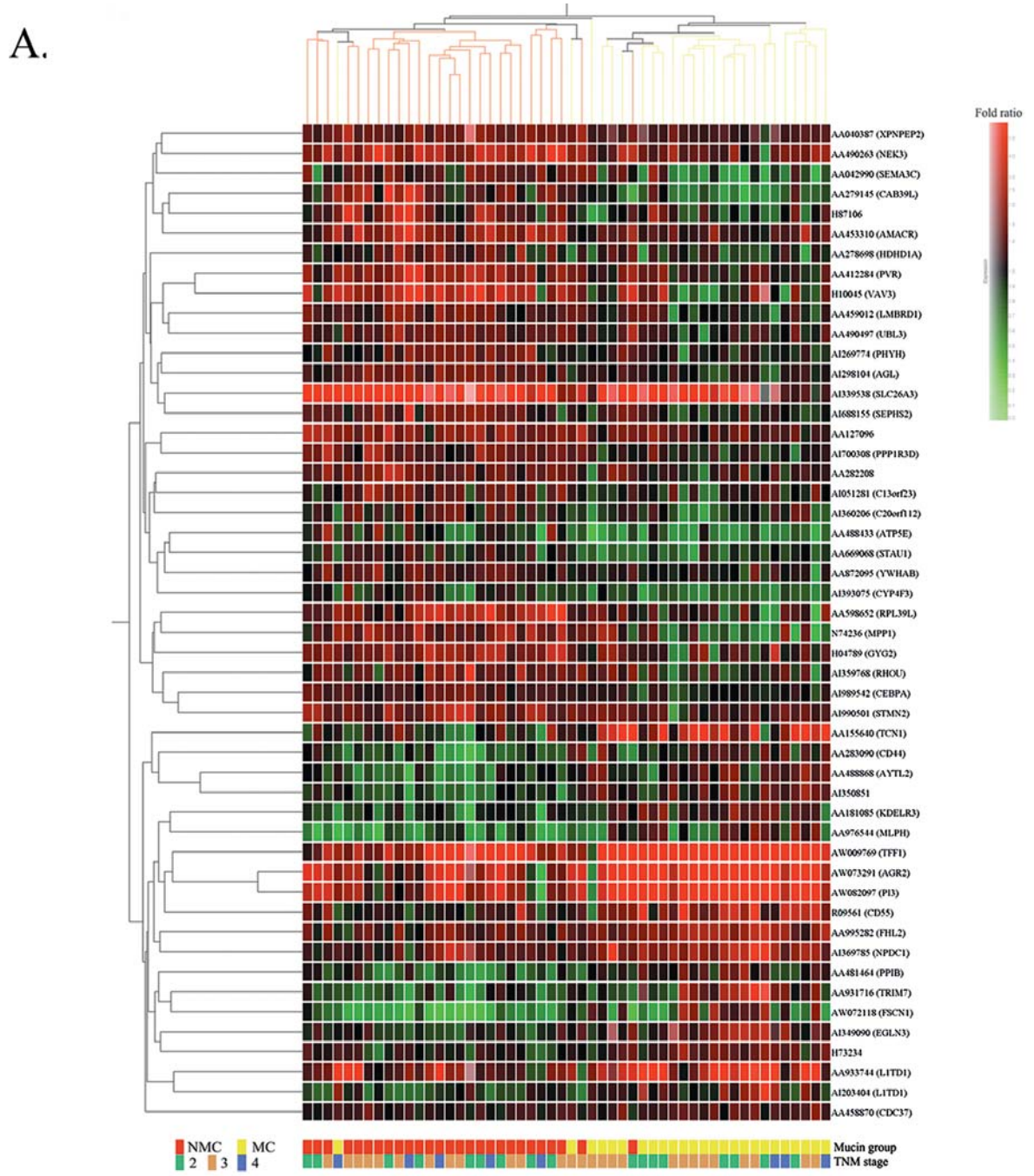

Figure 2. (A) The supervised clustering of training samples using 50 classifiers.

we selected 27 NMC samples with similar clinicopathologic factors to those of 25 MC samples including age, gender, tumor site, primary location, tumor depth and TNM stage (Table I). Between the two groups, these factors were balanced and validated by the χ^2 test.

BRAF and *KRAS* mutations were examined in 49 (24 MCs with 25 NMCs) and 42 samples (23 MCs with 19 NMCs), respectively. Two of 49 (4%) had a BRAF mutation, and both were MC. Twelve of 42 (29%) had a KRAS mutation, and 5 were MC (5 of 23, 22%) and 7 were NMC (7 of 19, 37%). Although not statistically significant, there were more *BRAF* mutations and fewer *KRAS* mutations in MC samples compared to NMC samples.

The classification analysis of mucinous and non-mucinous carcinomas. To find DEGs that can be used to discriminate between MC and NMC samples, we performed a Welch's t-test with the training set and 69 probes were selected with a threshold of false discovery rate of 0.05. After the annotation of 69 probes, 62 genes were determined to be DEGs and a classifier set (Table II). We validated these 62 genes using two statistical methods, the prediction analysis for microarrays (PAM) and support vector machines (SVMs), which are used for classification of samples with expression data (Table III). When using the gene signature, the prediction accuracy was 79.3 and 75.9% for PAM and SVM, respectively. Moreover, when the top 50 of 62 genes were used, the classification rate was 86.2 and 79.3% for PAM and SAM analysis,

B.

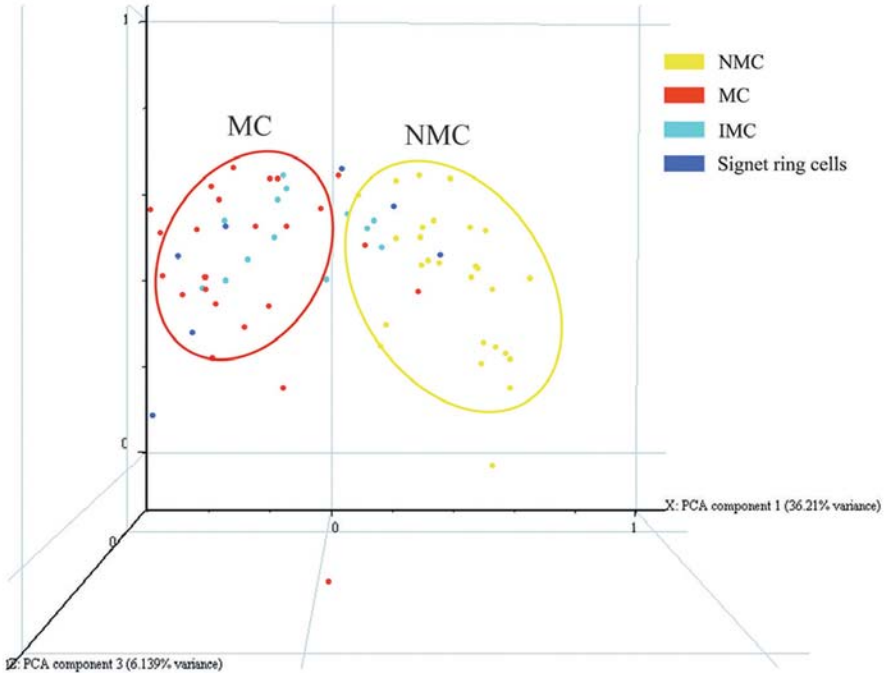

Figure 2. (B) Principal component analysis (PCA) of training samples, IMC and signet ring-cell carcinomas. The first principal component explains the differences in mucin volumes between samples. The x-, y- and z-axes represent the first three principal components, respectively.

respectively, suggesting that the 50-gene signature provided better prediction accuracy.

In the two-way hierarchical clustering and principal component analysis (PCA), the selected 50-gene signature clearly discriminated samples into NMCs and MCs (Fig. 2). However, in the case of intermediate mucinous carcinomas (IMC) and signet-ring cell carcinomas, samples could not be clearly classified based on the mucin contents.

Validation of selected classifiers. To investigate whether the 50-gene set could provide robust prediction for a larger data set, we performed LOOCV for all samples including the training, test, and validation sets. In this analysis, 81 out of 90 samples were correctly predicted (90%), showing better results in a large sample set than in an individual sample group. Finally, a 62-gene set that showed significant differential expression was selected and then the 50-gene set was shown to successfully discriminate MC and NMC.

The prediction of biological roles of selected DEGs. To functionally characterize the identified 62-gene set, the biological roles were investigated using Core Analysis in IPA 5.5 and various biological functions were predicted as major candidate functions with significant p-values. The top functions in the selected candidates were related to important cancer biological functions, including tumorigenesis (14 genes), cell cycle progression (6 genes), invasion (2 genes), anti-apoptosis (7 genes), and cell adhesion and proliferation (5 genes) together with carbohydrate metabolism (Table IV). In addition, many genes belonging to these functions were high-ranked molecules in Table III and almost all genes were novel molecules that were not known to have a relationship to MC. However, expression alterations in these genes in MCs had a positive relationship with tumorigenesis, cell cycle, invasion, anti-apoptosis and proliferation, suggesting the possibility that MCs have more severe clinical characteristics.

Quantitative real-time RT-PCR for the validation of microarray results. To validate the microarray results, we selected two highly up-regulated genes, *TFF1* and *AGR2*, and one significantly down-regulated gene, SLC26A3, and performed qRT-PCR to compare expression between NMCs and MCs. In addition, we also analyzed *MUC2*, which is known to be up-regulated in MCs and to be expressed at a 3.02-fold higher level between MCs and NMCs based on microarray results. The first 3 genes were shown to be significantly up- or down-regulated in MCs when compared to NMCs (Fig. 3A, $P<0.05$). MUC2 was significantly up-regulated by 6.33-fold in MCs compared to NMCs ($P<0.005$). For all 4 genes, changes in expression observed by qRT-PCR showed concordant results with the array data (Fig. 3B).

Discussion

Until now, most research on MC has focused on differences in clinical factors, variation of individual gene expression, and microsatellite instability. Frequent *BRAF* and infrequent *KRAS* mutations have been found in MC compared to NMC (8-10). Though more *BRAF* and fewer *KRAS* mutations were also observed in our study, the incidence of mutations was not statistically significant. With regard to microsatellite instability (MSI) status, MC (36%) showed a higher incidence

Table IV. Predicted biological functions and related genes of DEGs.

Category	Symbol	Description	Fold change[a]	FDR[b]
Cancer				
Tumorigenesis	AGR2	Anterior gradient homolog 2 (*Xenopus laevis*)	1.71	0.005
($P=1.07x10^{-12}$)	AMACR	α-methylacyl-CoA racemase	-0.66	0.029
	CD44	CD44 molecule (Indian blood group)	0.60	0.034
	CEBPA	CCAAT/enhancer binding protein (C/EBP), α	-0.42	0.005
	FHL2	Four and a half LIM domains 2	0.53	0.042
	FSCN1	Fascin homolog 1, actin-bundling protein (*Strongylocentrotus purpuratus*)	1.15	0.003
	FUS	Fusion [involved in t(12;16) in malignant liposarcoma]	-0.49	0.034
	GYG2	Glycogenin 2	-0.67	0.039
	MLPH	Melanophilin	0.73	0.034
	SEMA3C	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	-0.82	0.025
	SLC26A3	Solute carrier family 26, member 3	-2.47	0.038
	STMN2	Stathmin-like 2	-0.54	0.034
	TFF1	Trefoil factor 1	2.00	0.018
	VAV3	Vav 3 oncogene	-1.27	0.005
Cell division process	AMACR	α-methylacyl-CoA racemase	-0.66	0.029
($P=8.73x10^{-5}$)	CD44	CD44 molecule (Indian blood group)	0.60	0.034
	CEBPA	CCAAT/enhancer binding protein (C/EBP), α	-0.42	0.005
	VAV3	Vav 3 oncogene	-1.27	0.005
G2/M phase	AMACR	α-methylacyl-CoA racemase	-0.66	0.029
($P=1.78x10^{-3}$)				
Invasion	CD44	CD44 molecule (Indian blood group)	0.61	0.034
($P=2.75x10^{-4}$)	TFF1	Trefoil factor 1	2.01	0.018
Cell cycle				
Cell stage	AMACR	α-methylacyl-CoA racemase	-0.66	0.029
($P=4.24x10^{-6}$)	CD44	CD44 molecule (Indian blood group)	0.61	0.034
	CEBPA	CCAAT/enhancer binding protein (C/EBP), α	-0.42	0.005
	PVR	Poliovirus receptor	-0.81	0.003
	RHOU	Ras homolog gene family, member U	-0.64	0.034
	VAV3	Vav 3 oncogene	-1.27	0.005
Cell death				
Apoptosis	CD44	CD44 molecule (Indian blood group)	0.61	0.034
($P=7.22x10^{-5}$)	CD55	CD55 molecule, decay accelerating factor for complement (Cromer blood group)	0.82	0.046
	CEBPA	CCAAT/enhancer binding protein (C/EBP), α	-0.42	0.005
	FHL2	Four and a half LIM domains 2	0.53	0.042
	FUS	Fusion [involved in t(12;16) in malignant liposarcoma]	-0.49	0.034
	YWHAB	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, ß polypeptide	-0.49	0.005
	TFF1	Trefoil factor 1	2.01	0.018
Cell to cell signaling and interaction				
Adhesion	CD44	CD44 molecule (Indian blood group)	0.61	0.034
($P=1.07x10^{-4}$)	PPIB	Peptidylprolyl isomerase B (cyclophilin B)	0.62	0.005
	PVR	Poliovirus receptor	-0.81	0.003
	VAV3	Vav 3 oncogene	-1.27	0.005

 Continued.

Category	Symbol	Description	Fold change[a]	FDR[b]
Cellular growth and proliferation				
Proliferation	CD44	CD44 molecule (Indian blood group)	0.61	0.034
($P=2.39x10^{-5}$)	FSCN1	Fascin homolog 1, actin-bundling protein (*Strongylocentrotus purpuratus*)	1.15	0.003
	PVR	Poliovirus receptor	-0.81	0.003
Carbohydrate metabolism				
Metabolism of carbohydrate	CD44	CD44 molecule (Indian blood group)	0.61	0.034
($P=4.97x10^{-4}$)	GYG2	Glycogenin 2	-0.67	0.039
	TFF1	Trefoil factor 1	2.01	0.018

[a]All fold changes are the base 2 logarithms of fold ratios (MCs/NMCs). [b]FDR, false discovery rate.

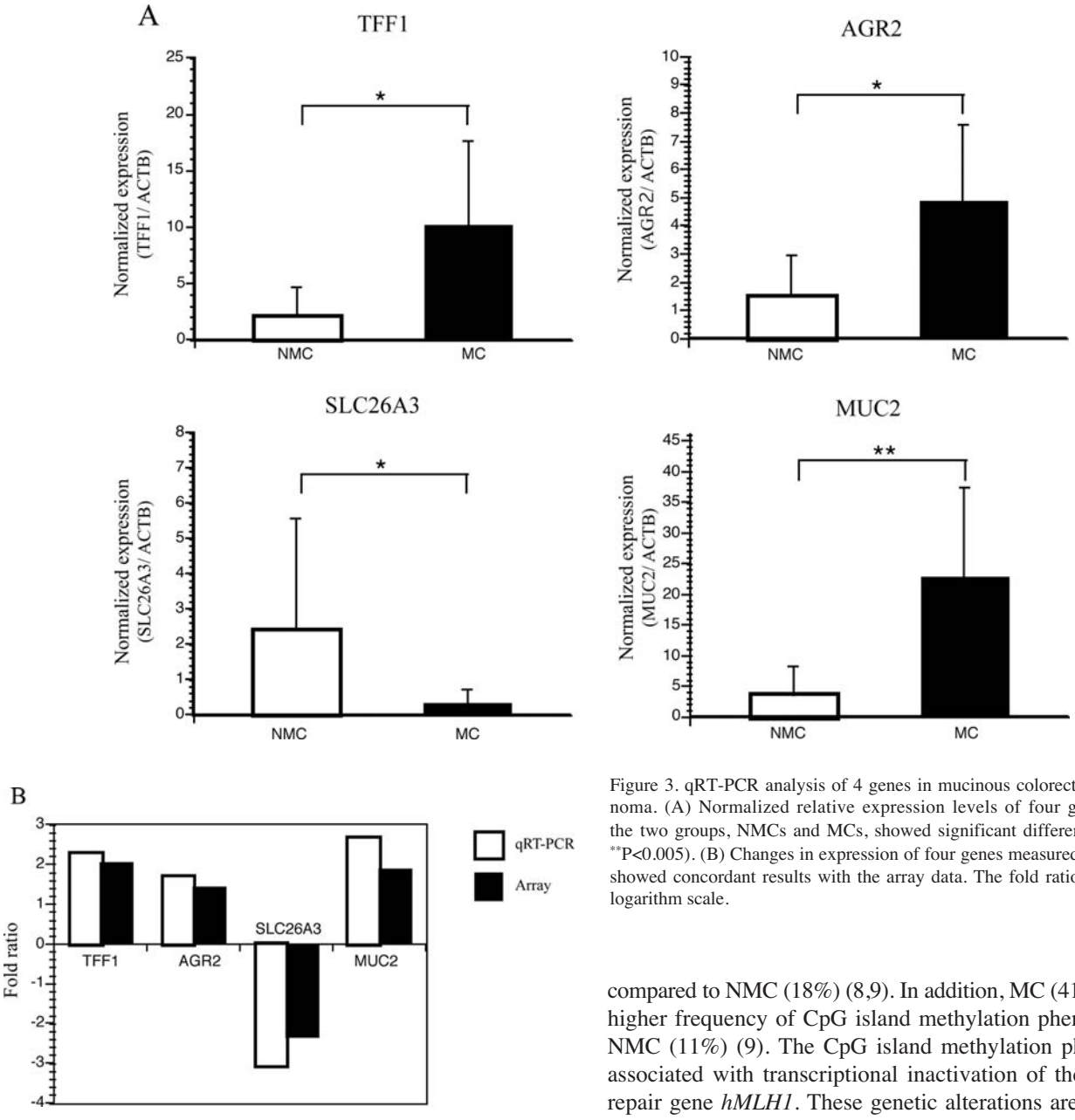

Figure 3. qRT-PCR analysis of 4 genes in mucinous colorectal adenocarcinoma. (A) Normalized relative expression levels of four genes between the two groups, NMCs and MCs, showed significant differences (*P<0.05, **P<0.005). (B) Changes in expression of four genes measured by qRT-PCR showed concordant results with the array data. The fold ratio is the base-2 logarithm scale.

compared to NMC (18%) (8,9). In addition, MC (41%) showed higher frequency of CpG island methylation phenotype than NMC (11%) (9). The CpG island methylation phenotype is associated with transcriptional inactivation of the mismatch repair gene *hMLH1*. These genetic alterations are thought to

be related to the molecular pathways underlying the development of MC. Though MC is thought to be associated with worse prognoses (4,5), biological studies have been limited to several molecules (6-12) and investigating the role of multiple genes is required to understand the nature of MC. Therefore, we preformed gene expression profiling of MC, making this the first study to use high-density cDNA microarray as an approach to understanding the biology of MC.

To select the best classifiers, several factors should be considered including sampling for data analysis, accurate pathologic data, microarray chip quality and analysis method. Among these factors, sampling method was important in the initial data analysis process. We classified samples as three sets, the training set to determine classifiers discriminating MC and NMC, the test set to select more predictive classifiers from the training set result, and the validation set to evaluate the prediction accuracy of selected classifiers. As MC is generally about 5% of total CRC patients, there were not enough MC samples for three whole sets. Of the three sets, the sample size and balanced tumor types were more important in the training set, where the gene set was determined. For the test and validation sets, sufficient sample numbers were more important than the balance of sample types because this process aimed to evaluate the classifiers for predicting each samples correctly. Therefore, we composed three sets of MC and NMC samples as shown in Fig. 1. Finally, we identified 50 genes that could classify the two groups with the best accuracy. In prediction and hierarchical clustering results, MC and NMC samples were clearly classified. However, intermediate mucinous carcinoma (IMC) and signet-ring cell carcinoma were not clearly classified based on their mucin component. These results might have arisen from a narrow distribution of mucin components and heterogeneous molecular characteristics.

To understand the molecular mechanisms underlying differences in clinical characteristics, we assessed molecular signatures of MC and NMC using significant DEGs. As a result, several biological functions implicated in cancer development and progression were selected (Table IV). These functions were tumorigenesis, cell cycle and cell proliferation, and almost all genes were novel molecules that had not previously been reported to be involved in MC. Among these genes, several genes were known to have a role associated with colorectal carcinoma. *SLC26A3* is known to be transcriptionally down-regulated in the early neoplastic process and underexpressed or absent in colon cancer samples compared with their normal counterparts (17-19). SLC26A3 dramatically suppressed colony formation and cell growth of various cancer cell lines, including colorectal cancer cells (20). In our results, *SLC26A3* was significantly down-regulated in MCs and down-regulation was also clearly confirmed in the quantitative real-time RT-PCR analysis (Fig. 3), which suggests that *SLC26A3* is involved in the carcinogenesis of MCs. FSCN1 increased invasiveness and proliferation of colon epithelial cell lines (21). Up-regulation of *FSCN1* was also associated with ER-negative breast carcinoma and a more aggressive clinical outcome (22). In our results, *FSCN1* was more up-regulated in MCs than in NMCs, which supports the worse prognosis with MC.

CD44 and TFF1 are key molecules with multiple functions. CD44 is a well-known transmembrane glycoprotein that plays a critical role in a variety of cellular behavior, including adhesion, migration, invasion and survival. In a previous report, it was reported that activation of CD44 decreases apoptosis of colon cancer cell lines (23) and increases adhesion of colorectal carcinoma cell lines and endothelial cell lines (24). In our results, expression of CD44 had an effect on tumor survival and proliferation. TFF1 is a secretory protein expressed in gastrointestinal mucosa. While the function of this gene is not well understood, it is thought to protect the mucosa from insults, stabilize the mucus layer, and affect healing of the epithelium. It has been reported that TFF1 could have important roles in apoptosis, tumorigenesis, invasion, and migration of colorectal cancer. TFF1 decreased apoptosis of colon cells (25), and increased invasiveness (26,27). Especially, TFF1 was expressed more intensively in stage A and B compared to C and D colorectal cancer tissues and was highly expressed in metastasized liver tissue (26). Immunohistochemistry studies revealed MC-specific staining of TFF1 (12). In our results, TFF1 was up-regulated in MCs compared to NMCs and in stage II compared to III and IV by 1.5-fold (data not shown). Further, in quantitative real-time RT-PCR, TFF1 was significantly up-regulated in MCs compared to NMCs (Fig. 3). In conclusion, TFF1 may play important roles in tumorigenesis, anti-apoptosis and invasiveness of MCs.

In our study, through the use of Core Analysis in IPA, three top ranked networks with selected DEGs were determined together with the biological roles, including cellular movement, cell-to-cell signaling and interaction and cell cycle (data not shown). This analysis provided valuable evidence to support the distinct molecular characteristics of MC, which will ultimately improve our understanding of the carcinogenesis of MC.

Using microarray analysis, we identified 50 classifiers out of 62 differentially expressed genes in mucinous carcinoma compared to non-MC. These genes were associated with tumorigenesis, invasion, cell cycle progression, anti-apoptosis, cell adhesion and proliferation. Based on the results of this study, we suggest that MCs have distinct molecular characteristics from NMC, and that the selected gene set provides a basis for a better understanding of the molecular pathogenesis of mucinous colorectal adenocarcinoma.

Acknowledgements

This study was supported by a Grant of the Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (0405-BC01-0604-0002).

References

1. Adell R, Marcote E, Segarra MA, Pellicer V, Gamon R, Bayon AM, Canales M and Torner A: Is mucinous colorectal adenocarcinoma a distinct entity? Gastroenterol Hepatol 25: 534-540, 2002.
2. Okuno M, Ikehara T, Nagayama M, Kato Y, Yui S and Umeyama K: Mucinous colorectal carcinoma: clinical pathology and prognosis. Am Surg 54: 681-685, 1988.
3. Nozoe T, Anai H, Nasu S and Sugimachi K: Clinicopathological characteristics of mucinous carcinoma of the colon and rectum. J Surg Oncol 75: 103-107, 2000.

SPANDIDOS PUBLICATIONS rti F, Lorenzotti A, Midiri G and Di Paola M: Prognostic cance of mucinous carcinoma of colon and rectum: a prospective case-control study. J Surg Oncol 73: 70-74, 2000.
5. Secco GB, Fardelli R, Campora E, Lapertosa G, Gentile R, Zoli S and Prior C: Primary mucinous adenocarcinomas and signet-ring cell carcinomas of colon and rectum. Oncology 51: 30-34, 1994.
6. Arai T, Kasahara I, Sawabe M, Kanazawa N, Kuroiwa K, Honma N, Aida J and Takubo K: Microsatellite-unstable mucinous colorectal carcinoma occurring in the elderly: comparison with medullary type poorly differentiated adenocarcinoma. Pathol Int 57: 205-212, 2007.
7. Kondo T, Masuda H, Abe Y and Takayama T: Two subtypes in colorectal mucinous carcinoma in relation to microsatellite instability. Hepatogastroenterology 49: 660-663, 2002.
8. Ogino S, Brahmandam M, Cantor M, Namgyal C, Kawasaki T, Kirkner G, Meyerhardt JA, Loda M and Fuchs CS: Distinct molecular features of colorectal carcinoma with signet ring cell component and colorectal carcinoma with mucinous component. Mod Pathol 19: 59-68, 2006.
9. Tanaka H, Deng G, Matsuzaki K, Kakar S, Kim GE, Miura S, Sleisenger MH and Kim YS: BRAF mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer. Int J Cancer 118: 2765-2771, 2006.
10. Yoshitake N, Fujii S, Mukawa K, Tominaga K, Fukui H, Ichikawa K, Tomita S, Ono Y, Imai Y, Terano A, Hiraishi H and Fujimori T: Mutational analysis of the BRAF gene in colorectal mucinous carcinoma in association with histological configuration. Oncol Rep 17: 9-15, 2007.
11. Ishizu H, Kumagai J, Eishi Y, Takizawa T and Koike M: Mucin core protein expression by colorectal mucinous carcinomas with or without mucus hyperplasia. J Gastroenterol 39: 125-132, 2004.
12. Kim DH, Kim JW, Cho JH, Baek SH, Kakar S, Kim GE, Sleisenger MH and Kim YS: Expression of mucin core proteins, trefoil factors, APC and p21 in subsets of colorectal polyps and cancers suggests a distinct pathway of pathogenesis of mucinous carcinoma of the colorectum. Int J Oncol 27: 957-964, 2005.
13. Clarke PA, te Poele R, Wooster R and Workman P: Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. Biochem Pharmacol 62: 1311-1136, 2001.
14. Wong YF, Selvanayagam ZE, Wei N, Porter J, Vittal R, Hu R, Lin Y, Liao J, Shih JW, Cheung TH, Lo KW, Yim SF, Yip SK, Ngong DT, Siu N, Chan LK, Chan CS, Kong T, Kutlina E, McKinnon RD, Denhardt DT, Chin KV and Chung TK: Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray. Clin Cancer Res 9: 5486-5492, 2003.
15. Kim TM, Jeong HJ, Seo MY, Kim SC, Cho G, Park CH, Kim TS, Park KH, Chung HC and Rha SY: Determination of genes related to gastrointestinal tract origin cancer cells using a cDNA microarray. Clin Cancer Res 11: 79-86, 2005.
16. Samowitz WS, Sweeney C, Herrick J, Albertsen H, Levin TR, Murtaugh MA, Wolff RK and Slattery ML: Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. Cancer Res 65: 6063-6069, 2005.
17. Antalis TM, Reeder JA, Gotley DC, Byeon MK, Walsh MD, Henderson KW, Papas TS and Schweinfest CW: Down-regulation of the down-regulated in adenoma (DRA) gene correlates with colon tumor progression. Clin Cancer Res 4: 1857-1863, 1998.
18. Byeon MK, Westerman MA, Maroulakou IG, Henderson KW, Suster S, Zhang XK, Papas TS, Vesely J, Willingham MC, Green JE and Schweinfest CW: The down-regulated in adenoma (DRA) gene encodes an intestine-specific membrane glycoprotein. Oncogene 12: 387-396, 1996.
19. Schweinfest CW, Henderson KW, Suster S, Kondoh N and Papas TS: Identification of a colon mucosa gene that is down-regulated in colon adenomas and adenocarcinomas. Proc Natl Acad Sci USA 90: 4166-4170, 1993.
20. Chapman JM, Knoepp SM, Byeon MK, Henderson KW and Schweinfest CW: The colon anion transporter, down-regulated in adenoma, induces growth suppression that is abrogated by E1A. Cancer Res 62: 5083-5088, 2002.
21. Jawhari AU, Buda A, Jenkins M, Shehzad K, Sarraf C, Noda M, Farthing MJ, Pignatelli M and Adams JC: Fascin, an actin-bundling protein, modulates colonic epithelial cell invasiveness and differentiation in vitro. Am J Pathol 162: 69-80, 2003.
22. Yoder BJ, Tso E, Skacel M, Pettay J, Tarr S, Budd T, Tubbs RR, Adams JC and Hicks DG: The expression of fascin, an actin-bundling motility protein, correlates with hormone receptor-negative breast cancer and a more aggressive clinical course. Clin Cancer Res 11: 186-192, 2005.
23. Bates RC, Edwards NS, Burns GF and Fisher DE: A CD44 survival pathway triggers chemoresistance via lyn kinase and phosphoinositide 3-kinase/Akt in colon carcinoma cells. Cancer Res 61: 5275-5283, 2001.
24. Fujisaki T, Tanaka Y, Fujii K, Mine S, Saito K, Yamada S, Yamashita U, Irimura T and Eto S: CD44 stimulation induces integrin-mediated adhesion of colon cancer cell lines to endothelial cells by up-regulation of integrins and c-Met and activation of integrins. Cancer Res 59: 4427-4434, 1999.
25. Bossenmeyer-Pourie C, Kannan R, Ribieras S, Wendling C, Stoll I, Thim L, Tomasetto C and Rio MC: The trefoil factor 1 participates in gastrointestinal cell differentiation by delaying G1-S phase transition and reducing apoptosis. J Cell Biol 157: 761-770, 2002.
26. Emami S, Le Floch N, Bruyneel E, Thim L, May F, Westley B, Rio M, Mareel M and Gespach C: Induction of scattering and cellular invasion by trefoil peptides in src- and RhoA-transformed kidney and colonic epithelial cells. FASEB J 15: 351-361, 2001.
27. Rodrigues S, Nguyen QD, Faivre S, Bruyneel E, Thim L, Westley B, May F, Flatau G, Mareel M, Gespach C and Emami S: Activation of cellular invasion by trefoil peptides and src is mediated by cyclooxygenase- and thromboxane A2 receptor-dependent signaling pathways. FASEB J 15: 1517-1528, 2001.