Cancer Informatics

ORIGINAL RESEARCH

Clinical Omics Analysis of Colorectal Cancer Incorporating Copy Number Aberrations and Gene Expression Data

Tsuyoshi Yoshida[1,2], Takumi Kobayashi[3], Masaya Itoda[1], Taika Muto[3], Ken Miyaguchi[1], Kaoru Mogushi[1], Satoshi Shoji[1], Kazuro Shimokawa[1], Satoru Iida[2], Hiroyuki Uetake[4], Toshiaki Ishikawa[4], Kenichi Sugihara[2], Hiroshi Mizushima[1,3] and Hiroshi Tanaka[1]

[1]Information Center for Medical Sciences, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan. [2]Department of Surgical Oncology, Graduate School of Medicine, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan. [3]Department of Medical Omics Informatics, School of Biomedical Science, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan. [4]Department of Translational Oncology, Graduate School of Medicine, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan. Corresponding author email: hmizushi@bioinfo.tmd.ac.jp

Abstract

Background: Colorectal cancer (CRC) is one of the most frequently occurring cancers in Japan, and thus a wide range of methods have been deployed to study the molecular mechanisms of CRC. In this study, we performed a comprehensive analysis of CRC, incorporating copy number aberration (CRC) and gene expression data. For the last four years, we have been collecting data from CRC cases and organizing the information as an "omics" study by integrating many kinds of analysis into a single comprehensive investigation.

In our previous studies, we had experienced difficulty in finding genes related to CRC, as we observed higher noise levels in the expression data than in the data for other cancers.

Because chromosomal aberrations are often observed in CRC, here, we have performed a combination of CNA analysis and expression analysis in order to identify some new genes responsible for CRC.

This study was performed as part of the Clinical Omics Database Project at Tokyo Medical and Dental University. The purpose of this study was to investigate the mechanism of genetic instability in CRC by this combination of expression analysis and CNA, and to establish a new method for the diagnosis and treatment of CRC.

Materials and methods: Comprehensive gene expression analysis was performed on 79 CRC cases using an Affymetrix Gene Chip, and comprehensive CNA analysis was performed using an Affymetrix DNA Sty array. To avoid the contamination of cancer tissue with normal cells, laser micro-dissection was performed before DNA/RNA extraction. Data analysis was performed using original software written in the R language.

Result: We observed a high percentage of CNA in colorectal cancer, including copy number gains at 7, 8q, 13 and 20q, and copy number losses at 8p, 17p and 18. Gene expression analysis provided many candidates for CRC-related genes, but their association with CRC did not reach the level of statistical significance. The combination of CNA and gene expression analysis, together with the clinical information, suggested UGT2B28, LOC440995, CXCL6, SULT1B1, RALBP1, TYMS, RAB12, RNMT, ARHGDIB, S1000A2, ABHD2, OIT3 and ABHD12 as genes that are possibly associated with CRC. Some of these genes have already been reported as being related to CRC. TYMS has been reported as being associated with resistance to the anti-cancer drug 5-fluorouracil, and we observed a copy number increase for this gene. RALBP1, ARHGDIB and S100A2 have been reported as oncogenes, and we observed copy number increases in each. ARHGDIB has been reported as a metastasis-related gene, and our data also showed copy number increases of this gene in cases with metastasis.

Conclusion: The combination of CNA analysis and gene expression analysis was a more effective method for finding genes associated with the clinicopathological classification of CRC than either analysis alone. Using this combination of methods, we were able to detect genes that have already been associated with CRC. We also identified additional candidate genes that may be new markers or targets for this form of cancer.

Keywords: colorectal cancer, clinical omics, microarray, copy number aberration

Cancer Informatics 2010:9 147–161

This article is available from http://www.la-press.com.

Background

Colorectal Cancer (CRC) has now become the third leading cause of death in Japan, and is one of the cancers with highest incidence in women.[1] In 2006, 22,547 men and 18,834 women died from CRC in Japan. Surgical operation is the most common treatment, but chemotherapy is also performed in Western countries. New drug targets have been developed recently, including oncogenes, anti-oncogenes, signal-transduction factors and apoptosis factors, and some of the drugs related to these factors are under clinical trial. Meanwhile, drug resistance is increasingly becoming an issue for chemotherapy.

Since 2005, we have been collecting comprehensive clinical information and omics information to establish our Integrated Clinical Omics Database (iCOD).[2,3] We have collected about 200 hepatocellular carcinoma cases, 200 CRC cases and 150 oral cancer cases. The database can be accessed at http://omics.tmd.ac.jp/. We are collecting comprehensive clinical information from patient records held at the hospital, and by interviews with patients. The data is anonymized and standardized for statistical analysis. The surgical specimens are stored, and genomic (single nucleotide polymorphisms (SNPs) and CNA), epigenomic, transcriptomic (mRNA and micro-RNA) and proteomic analysis are performed for each case. Our goal is to integrate comprehensive information to determine the relation between omics and clinical pathology by a systematic biomedical approach.

A lot of work has been done on molecular biological researches in CRC, and it is known that a specific sequence of several genes change sequentially during the development of CRC i.e. APC, beta-catenin, K-ras, p53, TDF-beta receptor, Smad2 and Smad4. This is called the adenoma-carcinoma sequence. A change in APC is observed in the early stage of this sequence, and thus APC is often called a gatekeeper gene for CRC.

CRC is considered to develop through two pathways: 85% of cases arise due to chromosomal instability (CIN), and 15% arise due to microsatellite instability or replication errors.[4–7] Copy number aberrations are often observed in CRC,[8–16] including gains at 7, 8q, 13 and 20q, and losses at 8p, 17p and 18. Some of these observations are related to the grade or metastasis level of CRC. Comprehensive analysis of these changes has previously been performed by comparative genomic hybridization or array comparative genomic hybridization, but a new method using SNP array has recently become possible.[17–20]

Samples

We obtained fresh-frozen tissue samples from 70 CRC patients (43 males and 27 females) who had undergone surgical resection at Tokyo Medical and Dental University Hospital Faculty of Medicine (Tokyo, Japan) between November 2005 and August 2007. This research project was reviewed and approved by institutional review board guidelines. Informed consent was obtained from all patients via the standard protocols of the institution. The tumors were located in the colon (46 samples) and rectum (24 samples). Normal tissue was obtained from each patient, taken from at the adjacent region of the colon or rectum, and used as the control sample. All resected specimens were collected in cryotubes, frozen and stored at −80 °C until the DNA and RNA analyses were performed.

DNA Isolation

Tumors were microdissected by removing the surrounding non-neoplastic tissue. Tumor DNA was extracted and purified using a QIAamp DNA Micro Kit (QIAGEN, Valencia, CA) according to the manufacturer's instructions. Contaminated RNAs were eliminated using RNase A during the purification process. Non-neoplastic tissues were homogenized in microtubes. Non-neoplastic tissue DNA was extracted and purified using a QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's instructions. Contaminated RNAs were eliminated using RNase A during the purification process. According to the Mapping 500 K Assay Manual supplied by Affymetrix (Santa Clara, CA), the minimum amount of genomic DNA required was 250 ng. Therefore, only purified DNA samples containing more than 250 ng of genomic DNA, as determined by a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE), were used in the subsequent experiments. For each sample, quality controls, which included confirming DNA degradation and contamination of RNAs, were performed by running 1% agarose gel electrophoresis with lambda DNA/Hind fragments.

SNP Array Analysis

The experiment was performed by strictly following the assay manual and using a GeneChip® Human Mapping 250 K Sty array (Affymetrix). The concentration of the starting materials was standardized by diluting genomic DNA samples to 50 ng/µl with a reduced EDTA-TE buffer. Since the Mapping 500 K Array set consists of two types of arrays, there are two alternatives of enzymes, which are Nsp I and Sty I. In this study, we used Sty I for digesting genomic DNA and used adaptor for Sty I in the PCR reaction following the manufacturers manual, which includes reaction at 94 °C 3 min, (94 degree 30 sec : 60 degree 45 sec : 68 degree 15 sec) × 30 cycles, 68 degree 7 min, and 4 degree hold. Amplicons were fragmented after the purification, followed by labeling reaction. After 16 hours of hybridization at 49 °C, the microarrays were transferred to a Fluidic Station 450 (Affymetrix), which is a totally automated system, for the washing and staining steps. After fluorescence staining, the microarray images were scanned by an Affymetrix laser scanner.

Data Analyses of SNP Arrays

The microarray data from the laser scanner was used for copy number analysis with a Chromosome Copy Number Analysis Tool (Affymetrix). The copy number for each SNP probe set taken from a tumor sample was calculated by comparing the probe intensity to the reference probe intensity from non-neoplastic tissue, and creating a list of the genome-wide copy number data. Genome-wide CNA analysis was carried out for the gene information combined with the copy numbers in order to identify the overall tendency of the chromosomal CNAs over the whole human genome in CRCs. Software written in the R language (http://www.r-project.org/) was used to perform this visualization of CNAs across the chromosomes and for the rest of the data analyses.

Total RNA Isolation

Tumors were microdissected by removing the surrounding non-neoplastic tissue. The total RNA was extracted and purified using an RNeasy micro kit (QIAGEN) with on-column DNase digestion, according to the manufacturer's instructions. The integrity of the total RNA we obtained was assessed using a Bioanalyzer RNA 6000 Nano Assay (Agilent Technologies, Palo Alto, CA). Samples with an RNA Integrity Number (RIN) greater than 5.0 were used for the rest of the experiments.

Microarray Analysis

Using 100 ng of total RNA, cRNA was prepared using two-cycle target labeling and control reagents, namely Affymetrix P/N 900494 (Affymetrix). The experiment was performed using a GeneChip® Human Genome U133 Plus 2.0 Array (Affymetrix) in strict adherence to the assay manual. The 70 output data files (CEL files) obtained by this process were then normalized with the robust multi-array average method using R 2.4.1 statistical software (http://www.r-project.org/) together with the "Affy" package from BioConductor (http://www.bioconductor.org/). Estimated gene expression levels were calculated as log2-transformed values, and 62 control probe sets were removed for further analysis. In order to identify differently expressed genes associated with clinico-pathological characteristics, we performed exact Wilcoxon rank-sum tests available from the in "Coin" package. Fold-change values (FC) were also calculated using the ratio of geometric means of gene expression levels in each patient group.

Gene Sets

Our CNA data was also statistically analyzed to investigate the association between the CNA and the prognosis of CRC. Fisher's exact test was applied to multiple conditions, such as T/N/M classification and recurrence. All genes with P-values < 0.01 were classified into Gene Set I.

Statistically significant differences in gene expression were assessed using the Wilcoxon exact rank-sum test from the exactRankTests package. The Wilcoxon exact rank-sum test was also applied to multiple conditions, such as T/N/M classification and recurrence. Genes with P-values < 0.05 and FC > 1.5 were classified into Gene Set II.

Genes included in both Gene Set I and Gene Set II were defined as Gene Set III.

Results

Analysis of copy number variations

Based on the copy number variation analysis of 70 cases, we found copy number increases in chromosome 7, 8q, 13 and 20q, and copy number decreases in 8p, 17p and 18 (Fig. 1). In particular, copy number increases in chromosome 20q and decreases in chromosome 18q were identified in over 80% of cases. Gene amplifications were also involved in these chromosomes: for example, the CRC-associated gene snail homolog 1 (SNAI1) exists in chromosome 20q and is deleted in colorectal carcinoma (DCC) in chromosome 18q.

Selection of Gene Set I (from the relationship between copy number variation and clinico-pathological factors)

We performed Fisher's exact tests in order to examine the significance of any associations between the copy numbers of all genes (26,376 genes) and clinico-pathological factors, and selected the statistically significant genes ($P < 0.01$) (Gene Set I). These genes were used in subsequent analyses for selecting Gene Set III.

In one association analysis between T classification (T1 and T2 vs. T3 and T4) and copy number variations, three genes showed significant copy number increases and 28 genes showed significant copy number decreases. These genes were categorized as Gene Set I-T_A.

In another association analysis between T classification (T1, T2 and T3 vs. T4) and copy number variations, 407 genes showed significant copy number increases and 209 genes showed significant copy number decreases. These genes were categorized as Gene Set I-T_B.

In association analysis between N classification (N0 vs. N1 and N2) and copy number variations, 145 genes showed significant copy number increases and 88 genes showed significant copy number decreases. These genes were categorized as Gene Set I-N.

In association analysis between M classification (M0 vs. M1) and copy number variations, 170 genes showed significant copy number increases and 124 genes showed significant copy number decreases. These genes were categorized as Gene Set I-M.

In association analysis between recurrence (Re0: no recurrence vs. Re1: recurrence) and copy number variations, 391 genes showed significant copy number increases and 311 genes showed significant copy number decreases. These genes were categorized as Gene Set I-Re.

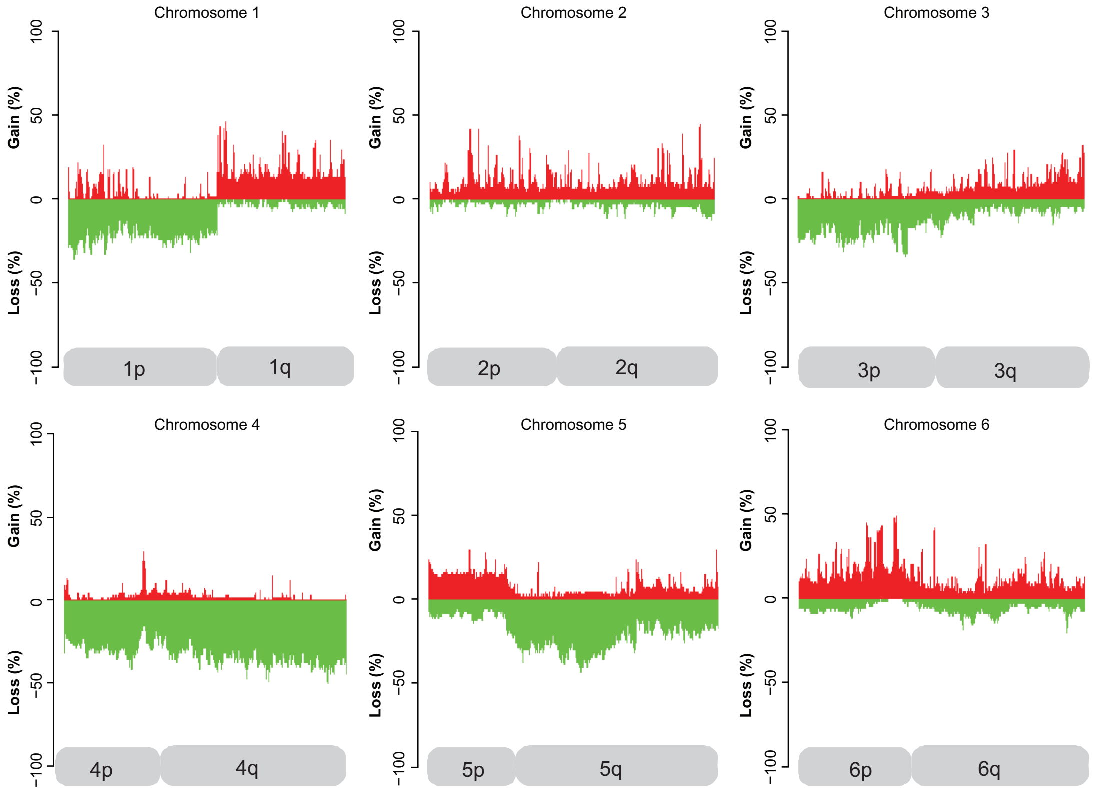

Figure 1. (*Continued*)

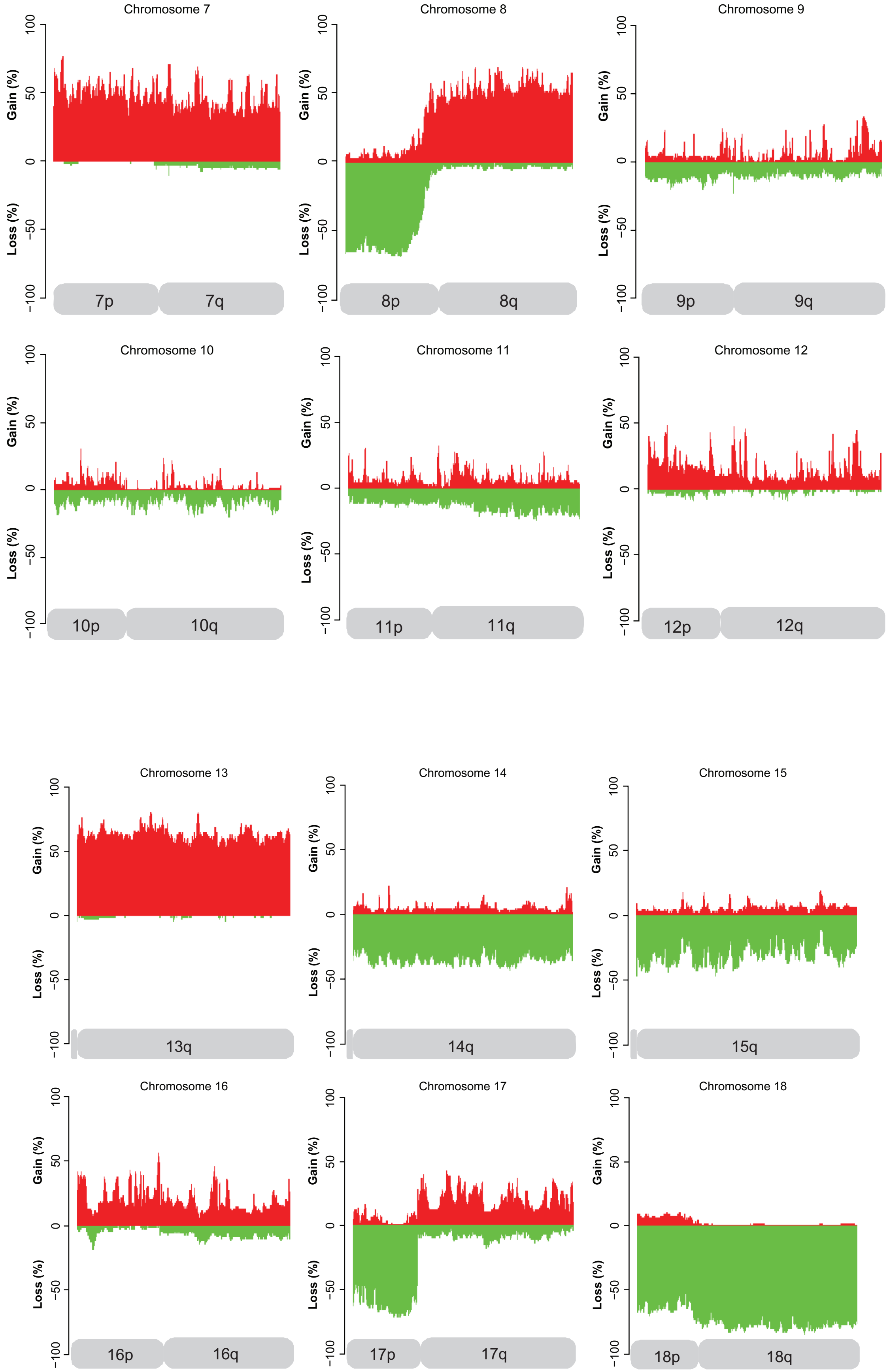

Figure 1. (*Continued*)

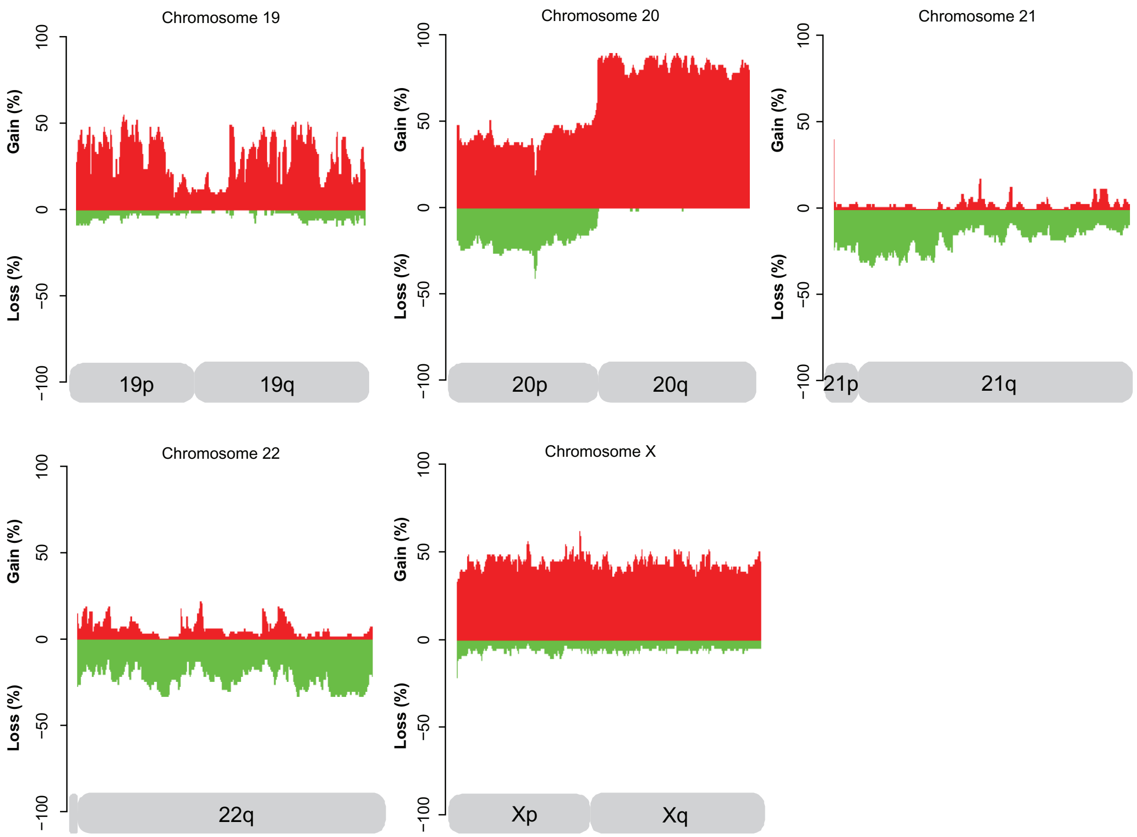

Figure 1. Percentage of CNA at each probe set position in 70 CRC cases sorted by chromosomes.
Note: Percentages of cases with a copy number aberration (red: gain; green: loss) at each probe set are shown as a bar-graph.

Selection of Gene Set II (from the relationship between gene expressions and clinico-pathological factors)

All 70 cases were classified into two groups by clinicopathological criteria, and Wilcoxon exact rank tests using the expression data of each gene were performed. We selected some statistically significant genes (P-value < 0.05 and FC > 1.5, or P-value < 0.05 and FC < 0.67) and analyzed these genes as Gene Set II.

As to T classification (T1 and T2 vs. T3 and T4), 29 genes were significantly up-regulated and 89 genes were significantly down-regulated. We categorized these genes as Gene Set II-T_A and used them in subsequent analyses.

As to T classification (T1, T2 and T3 vs. T4), 30 genes were significantly up-regulated and 52 genes were significantly down-regulated. We categorized these genes as Gene Set II-T_B.

As to N classification (N0 vs. N1 and N2), 39 genes were significantly up-regulated and 26 genes were significantly down-regulated. We categorized these genes as Gene Set II-N.

As to M classification (M0 vs. M1), 100 genes were significantly up-regulated and 46 genes were significantly down-regulated. We categorized these genes as Gene Set II-M.

As to recurrence classification (Re0 vs. Re1), 93 genes were significantly up-regulated and 54 genes were significantly down-regulated. We categorized these genes as Gene Set II-Re.

Selection of Gene Set III (from the relationship between copy number variations and gene expressions)

The genes that appeared in both Gene Set I and Gene Set II were classified as Gene Set III. Genes in this third set had parallel expression and copy number change (e.g. copy number increase paralleled up-regulated expression). The subcategorirs in Gene Set III followed the same clinico-pathological labeling system as Gene Sets I and II. Figure 2 compares the expression differences for expression differences for each gene in Gene Set III and Table 2 lists the total number of genes in each set. Table 3 lists the genes

Table 1. Clinico-pathological data for 70 primary CRC cases.

Factor	Patients (*n* = 70)	Percentage (%)
Age (years)	65.7 ± 10.9 (range: 33–87)	
Gender		
Male	43	61.4
Female	27	38.6
Position		
Colon	46	65.7
Rectum	24	34.3
Stage		
1	9	12.9
2	25	35.7
3	23	32.9
4	13	18.6
T		
1	1	1.4
2	12	17.1
3	28	40.0
4	29	41.4
N		
0	39	55.7
1	23	32.9
2	8	11.4
M		
0	57	81.4
1	13	18.6
Recurrence		
Re–	50	71.4
Re+	20	28.6

appearing in Gene Set III by clinico-pathological classification.

In Gene Set III-T_A, no genes showed both a copy number increase and up-regulated expression (FC > 1.5 and $P < 0.05$; group 1); however, UDP glucuronosyltransferase 2 family polypeptide B28 (UGT2B28) showed both a copy number decrease and down-regulated expression (FC > 0.67 and $P < 0.05$; group 2).

In Gene Set III-T_B, group 1 contained a hypothetical gene supported by BC034933; BC068085 (LOC440995) was identified. Group 2 contained chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2) (CXCL6); the sulfotransferase family; cytosolic, 1B, member 1 (SULT1B1) and chemokine (C-X-C motif) ligand 3 (CXCL3).

In Gene Set III-N, group 1 contained ralA binding protein 1 (RALBP1); thymidylate synthetase (TYMS); RAB12, a member of the RAS oncogene family (RAB12); and RNA (guanine-7-) methyltransferase (RNMT); group 2 contained no genes.

In Gene Set III-M, group 1 contained no genes, and group 2 contained Rho GDP dissociation inhibitor (GDI) beta (ARHGDIB).

In Gene Set III-Re, group 1 contained S100 calcium binding protein A2 (S100A2) and abhydrolase domain containing 2 (ABHD2), and group 2 contained oncoprotein induced transcript 3 (OIT3) and abhydrolase domain containing 12 (ABHD12).

Thus a total of 14 genes were assigned to Gene Set III, which was a much smaller number of genes than assigned to Gene Set I or Gene Set II.

When we remove fold change from the analysis, the number of selected genes increases dramatically. As many of the genes showed no correlation between copy number and gene expression, we defined Gene Set III in terms of both FC and *P*-value.

Discussion

The results of the CNA analysis (Fig. 1) clearly show that copy number gains were often observed in chromosomes 7, 8q, 13 and 20q. Copy number losses were also frequently observed in chromosomes 8p, 17p, and 18. These results match the observations in other institutes and even in other countries,[8–16] so we believe these characteristics can be considered to apply to CRC worldwide, despite differences in genetic backgrounds or lifestyles.

We observed a copy number gain in SNAI1, a CRC-related gene, in more than 85% of cases, as a CRC-related gene. SNAI1 is located at 20q13.1-q13.2, and is said to play a role in the regulation of the cell adhesion protein E-cadherin (CDH1).[21] SNAI1 is also plays a role in phosphorylation through Axin and GSK-3beta for degradation of beta-catenin in proteasomes in the Wnt-signaling pathway.[22] DCC is located at 18q21.3, and it is one of the membrane proteins with netrin-1 (NTN1) as a ligand. DCC of SNAI1 makes an apoptotic signal to the cell when no ligand is bound, and NTN1 inhibits its signal, thus DCC and NTN1 are thought to be one of the cancer suppressor genes. Repression of DCC is observed not only in CRC, but also in other tumors.[23] The copy number changes observed in the present study also support these previous findings.

Many genes were extracted as those related to the clinico-pathological classifications using Fisher's test in Gene Set I ($P < 0.001$). However, most of them were unknown open reading frames or pseudogenes. Even it was difficult to establish an association between the annotated genes and CRC. A similar analysis was performed for Gene Set II using the Wilcoxon exact rank test for clinico-pathological classification, and some genes were extracted ($P < 0.05$) and FC > 1.5. However, this was insufficient to infer a relationship with CRC. For this reason, we performed a combined analysis of Gene Sets I and II.

We were able to extract several genes by this combined analysis, and defined these as Gene Set III. We investigated these genes further by their function and by a search of the literature.

Gene Set III-T_A

UGT2B28 was extracted as Gene Set III-T_A (T1 and T2 vs. T3 and T4), which showed copy number loss and a decrease in expression level (FC < 0.67, $P < 0.05$). UGT2B28 is located at 4q13.2, and is a membrane protein in cytosolic microsomes as one of the subtypes of uridine diphospho-glucuronosyl-transferases (UGTs). UGTs are enzymes for glucronization of foreign molecules for detoxification. UGT2B plays a role in the metabolism of bilic acid, all-trans retinoic acid (ATRA), non-steroidal anti-inflammatory drugs (NSAIDs), flavonoids and steroids.[24,25] Substrates of UGT2B generally have an inhibitory effect for CRC, but down-regulation of UGT2B28 may lead to an abnormal environment in the cell.

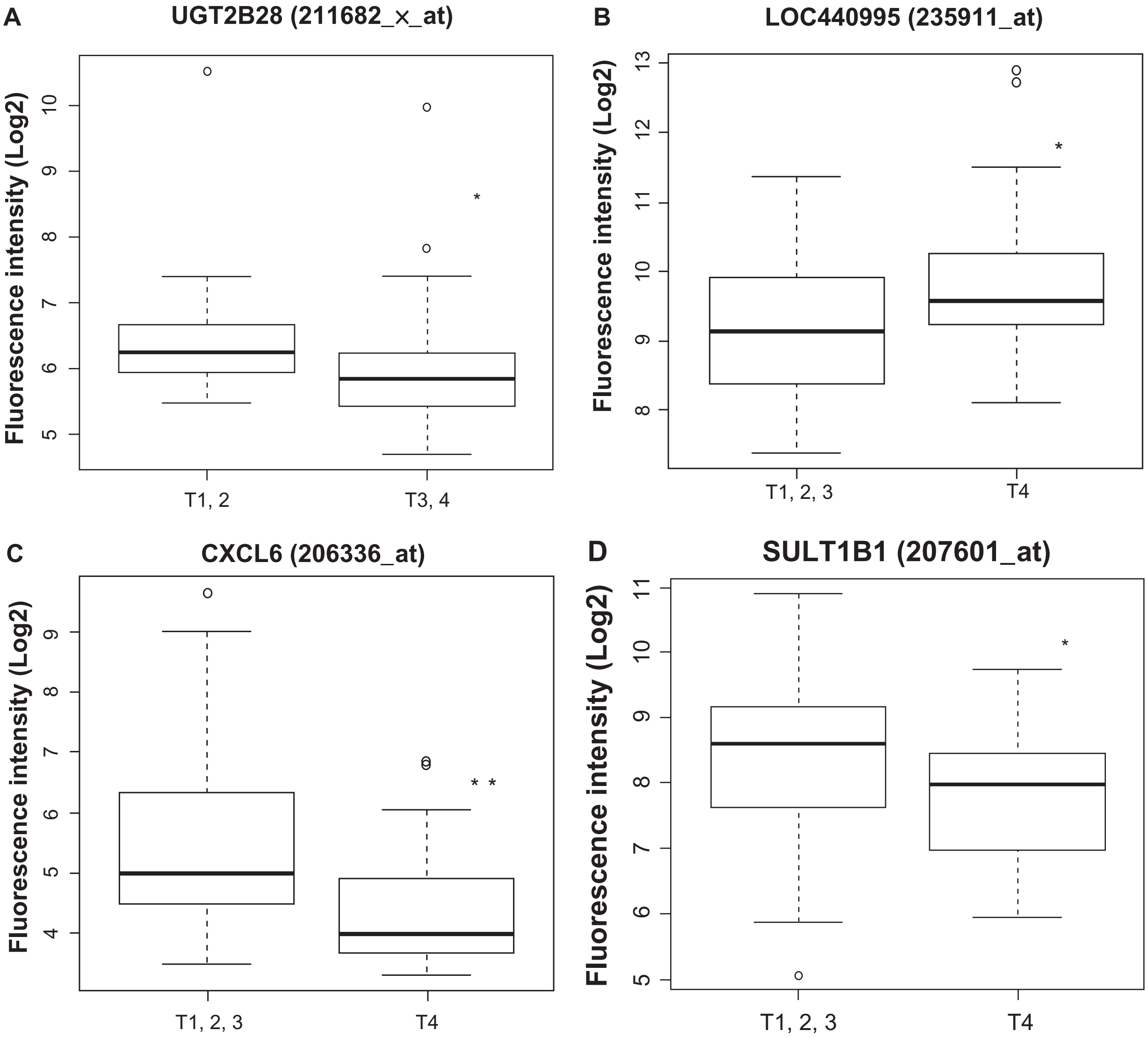

Figure 2. (*Continued*)

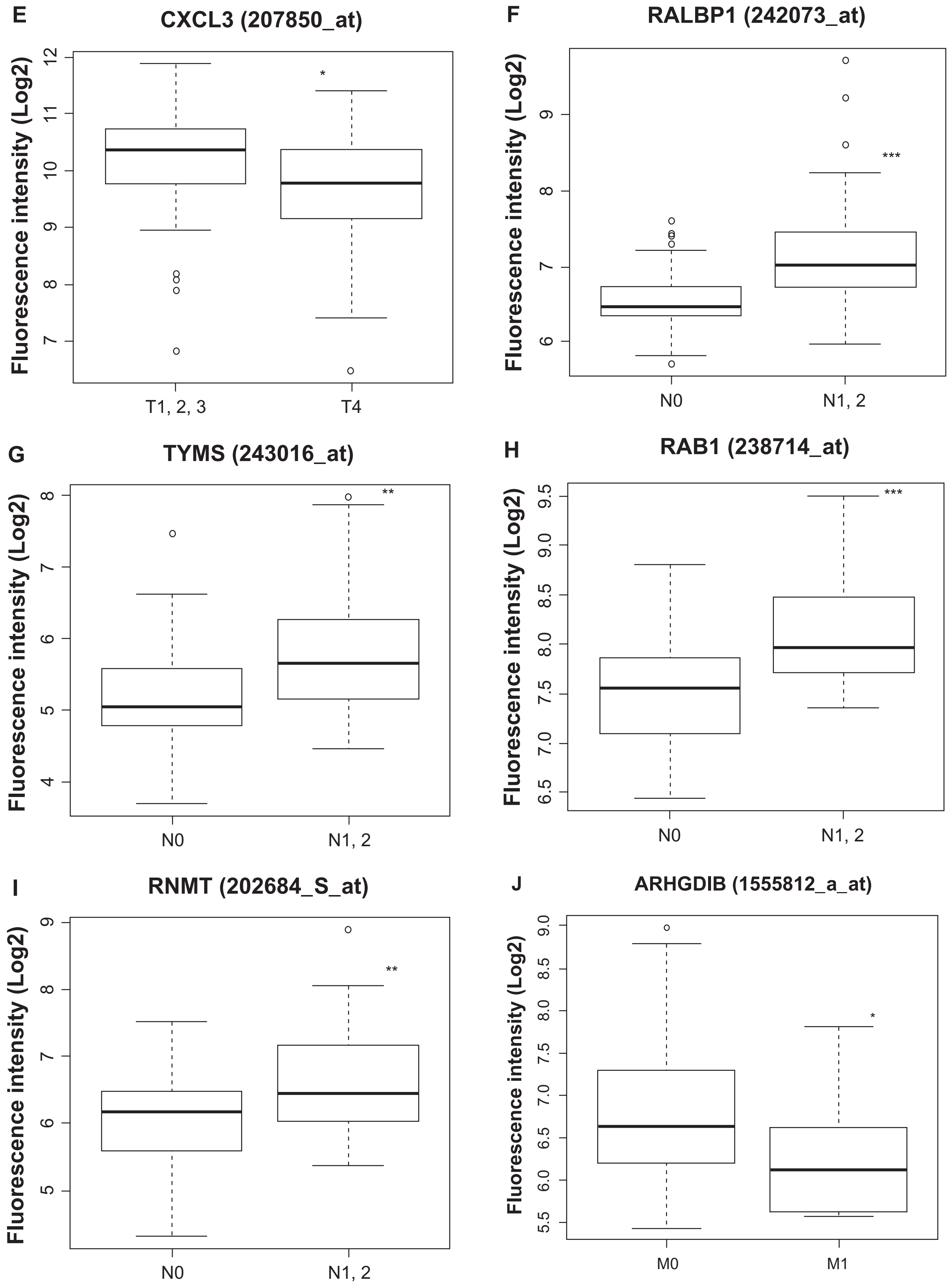

Figure 2. (*Continued*)

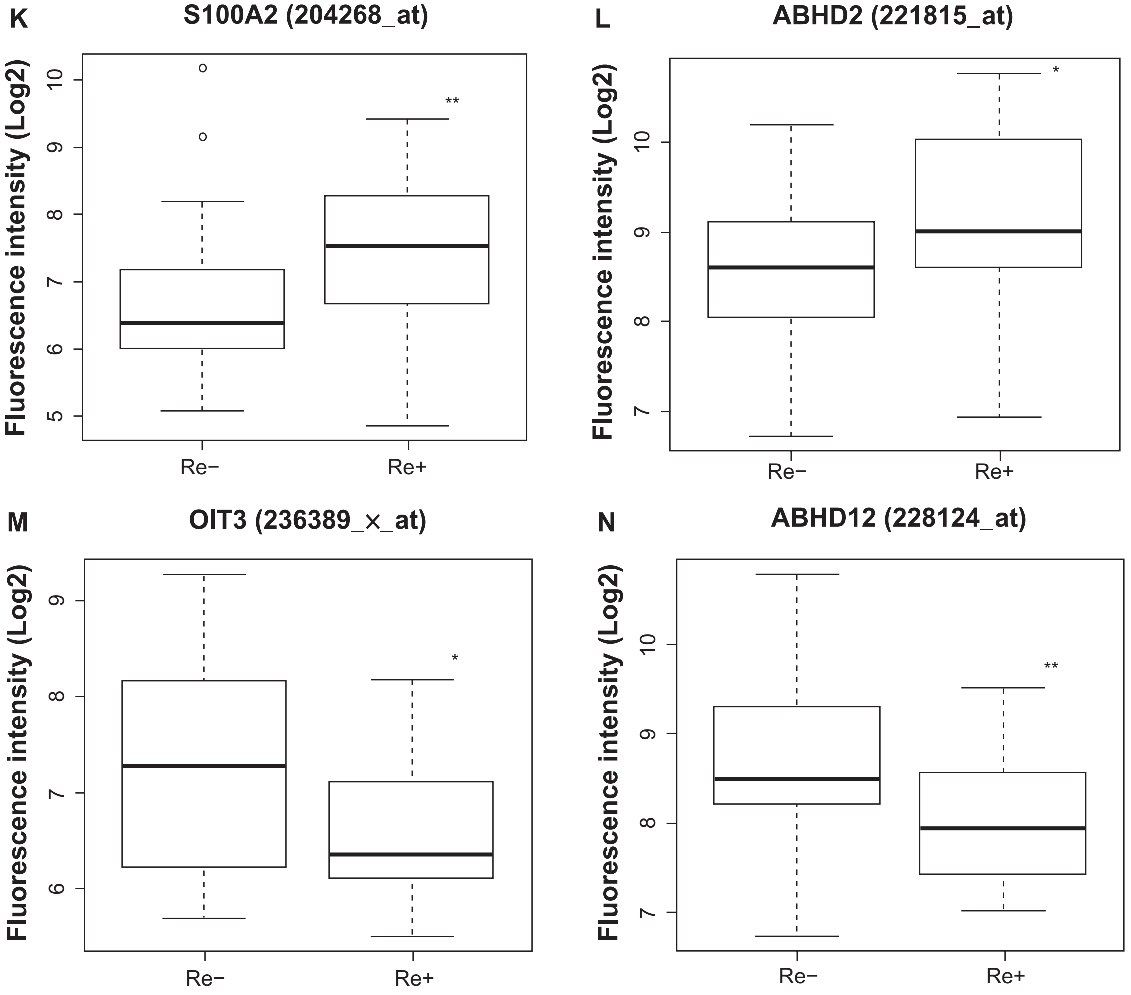

Figure 2. Expression comparison of each genes in Gene Set III. Expression differences are shown for each gene, compared by each group. Gene Set III-T_A (T1 and T2 vs. T3 and T4): **A**) UGT2B28, **B**) LOC440995, **C**) CXCL6, **D**) SULT1B1; Gene Set III-T_B (T1, T2 and T3 vs. T4): **E**) CXCL3, **F**) RALBP1, **G**) TYMS, **H**) RAB12; Gene Set III-N (N0 vs. N1 and N2): **I**) RNMT; Gene Set III-M (M0 vs. M1): **J**) ARHGDIB, **K**) S100A2, **L**) ABHD2, **M**) OIT1; Gene Set III-Re (recurrence vs. non-recurrence): **N**) ABHD12.
Notes: $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Gene Set III-T_B

LOC440995 was extracted as Gene Set III-T_B (T1, T2 and T3 vs. T4), which showed copy number gain and a high expression level (FC > 1.5 and $P < 0.05$). CXCL6, SULT1B1 and CXCL3 genes were also extracted in Gene Set III as copy number loss and expression level down (FC < 0.67 and $P < 0.05$). The LOC440995 gene is located at 3q29; however, neither its function nor annotation was clear.

The CXCL6 gene is located at 4q21, its product is one of the ligands for the chemokine receptor interleukin 8 receptor (IL8R) and it exhibits neutrophil migration activity. CXCL6 also has angiogenesis activity and can stimulate cancer cell progression.[26,27] Zhu et al reported overexpression of CXCL6 with its receptor, IL8R, in hypoxic small cell lung cancer, which is an autocrine regulation for cell progression.[28] Rubie et al observed an overexpression of CXCL1 and CXCL5 in CRC, but the level of CXCL6 was not changed.[29] Our observations showed that copy number loss and low expression levels were correlated with the invasiveness of CRC. To understand this relationship better, we must further clarify the function of CXCL6 in CRC.

The SULT1B1 gene is located at 4q13.3, and is one of the sulfotransferase family members (SULTs). SULT2 mainly contributes to the metabolism of steroids.[30] Sulfate conjugation is usually a kind of detoxification, and sometimes activates some chemical compound. For example, the carcinogens, 2-amino-1-methyl-6-phenylimidazo(4,5-b) pyridine and

Table 2. Numbers of genes in each gene set associated with key clinico-pathological factors.

Clinicopathological factor	Gene Set		
	Gene Set I (CNA)	Gene Set II (EC)	Gene Set III CNA + EC
T1 and T2 vs. T3 and T4	31	118	1
T1, T2 and T3 vs. T4	616	82	4
N0 vs. N1 and N2	233	65	4
M0 vs. M1	294	146	1
Re– vs. Re+	702	147	4
Total	1,876	558	14

7,12-dimethylbenz(a)anthracene gain carcinogenetic function by sulfate conjugation by SULT1. High expression of SULT1B1 mRNA has been reported in normal colorectal samples, and it is thought to be related to carcinogenesis.[30] Expression of SULT1B1 is a characteristic for normal colorectal tissue, so repression of SULT1B1 along with repression of UGT2B28 in CRC is thought to be related to tissue dedifferentiation.

The CXCL3 gene is located at 4q21, and its product is one of the ligands for a chemokine receptor, the interleukin 8 receptor (IL8R), similar to the case of CXCL6. Recently, overexpression of CXCL3 has been observed in many cancers, including esophagus cancer[31] and breast cancer,[32] and has been suggested as being related to cancer recurrence. However, Li et al reported that CXCL3 expression had no relation with remote metastasis of CRC,[33] which means CXCL3 may have a different function in CRC. To clarify this matter, we need to validate the CNA and expression in detail, as we have planned to do for CXCL6.

Gene Set III-N

RALBP1, TYMS, RAB12 and RNMT were extracted as Gene Set III-N (N0 vs. N1 and N2). These showed copy number gains and high expression levels (FC > 1.5 and $P < 0.05$).

RALBP1 is located at 18p11.3, and is known to play a role in anti-apoptosis function and protection from stress by glutathione conjugation. The expression level of RALBP1 was shown to be increased in many cancer cells by expression array analysis. In tumor cells, RALBP1 not only suppresses apoptosis, but is also related with transportation of chemical anti-cancer compounds. Singhal et al found that antisense RNA for RALBP1 in cancer cell lines or

Table 3. List of genes in Gene Set III using each clinico-pathological classification grouping.

NCBI gene ID	Gene symbol	Gene title	Location
Invasion (T1 and T2 vs. T3 and T4 : Gene Set III-T_A)			
54490	UGT2B28	UDP glucuronosyltransferase 2 family, polypeptide B28	4q13.2
Invasion (T1, T2 and T3 vs. T4 : Gene Set III-T_B)			
440995	LOC440995	hypothetical gene supported by BC034933; BC068085	3q29
6372	CXCL6	chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)	4q21
27284	SULT1B1	sulfotransferase family, cytosolic, 1B, member 1	4q13.3
2921	CXCL3	chemokine (C-X-C motif) ligand 3	4q21
Lymph node metastasis (N0 vs. N1 and N2 : Gene Set III-N)			
10928	RALBP1	ralA binding protein 1	18p11.3
7298	TYMS	thymidylate synthetase	18p11.32
201475	RAB12	RAB12, member RAS oncogene family	18p11.22
8731	RNMT	RNA (guanine-7-) methyltransferase	18p11.22–p11.23
Distant metastasis (M0 vs. M1 : Gene Set III-M)			
397	ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta	12p12.3
Recurrence (Gene Set III-Re)			
6273	S100A2	S100 calcium binding protein A2	1q21
11057	ABHD2	abhydrolase domain containing 2	15q26.1
170392	OIT3	oncoprotein induced transcript 3	10q22.1
26090	ABHD12	abhydrolase domain containing 12	20p11.21

cancer cells in mice suppressed cell growth, and these effects were much more pronounced when combined with other anti-cancer drugs.[34]

TYMS is located at 18p11.32, and is a target for the anti-cancer drug 5-fluorouracil (5-FU). This drug inhibits the reaction between deoxyuridine monophosphate (dUMP) and deoxythymidine monophosphate (dTMP), resulting in the inhibition of DNA synthesis. However, the expression of TYMS is up-regulated in most cancers, which, in turn, makes tumor cells resistant to 5-FU treatment. Sharma et al reported that 5-FU resistance was stronger in patients with overexpression of either TYMS mRNA or protein.[35] Although the relation between TYMS with lymph node metastasis is not clear, we are now analyzing its relationship with cancer resistance, as we observed the copy number gain and overexpression of TYMS more frequently in advanced CRC cases.

RAB12 is located at 18p11.22, and is known as one of the Rab GTP-binding proteins, which is a GTPase with low molecular weight. The Rab GTP-binding protein is associated with the membrane transporter, and takes part in the excretion of hormones or neuron messengers. Mosesson et al found that the integrin molecule is transported in the direction of cancer invasion by endocytosis of this molecule.[36] However, there has been no report of a direct relation between RAB12 and cancer. Iida reported that RAB12 is a vesicle-associated small GTPase that may be activate the transportation of the endoplasmic reticulum from cytosol to centrosome on the cytoskeleton.[37]

RNMT is located at 18p11.22-p11.23, and is known to play a role in the maturation of mRNA.[38] but there has been no report about its relation to cancer.

Gene Set III-M

ARHGDIB was extracted because of its copy number loss and low expression level (FC $<$ 0.67 and $P < 0.05$) in Gene Set III-M (M0 vs. M1). ARHGDIB is located at 12p12.3, and is known as a target gene of transcription factor v-ets erythroblastosis virus E26 oncogene homolog 1 (ETS1). Expression of ARHGCIB was found in hematopoietic cells and epithelial cancer. ETS1 is regulated by protein kinase C, alpha (PKCA) and is related to cancer invasion and progression. ARHGDIB, working with vav 1 guanine nucleotide exchange factor (VAV1), activates the nuclear factor of activated T-cells 1 (NFAT1) and results in transcription regulation of the tumor-related gene cyclooxygenase (COX2). COX2 is related to cancer progression, and Schunke reported that the activation of COX2 in breast cancer lowers the survival rate, and promotes metastasis to lung or bone.[39] On the other hand, ARHGDIB also has an ability to inhibit Rho GTPase, which inhibits cancer invasion. This relates to epithelial mesenchymal transformation. Rho GTPase inhibitors (Rho-GDI) such as ARHGDIB bind to GDP-bound Rho GTPase, and inhibits its function by stabilizing this bound form. These findings suggest that changes in the expression level of ARHGDIB result in a positive effect for remote metastasis of cancer.[39] Ota et al suggested that there was a relation between a low inhibitory effect of ARHGDIB and remote metastasis,[40] which supports our observations.

Gene Set III-Re

S100A2 and ABHD2 had copy number gains and high expression levels (FC $>$ 1.5 and $P < 0.05$) in Gene Set III-Re (No recurrence vs. recurrence). Within this same set, OIT3 and ABHD12 had copy number losses and low expression levels (FC $<$ 0.67 and $P < 0.05$).

S100A2 is one of the calcium-binding proteins and is located as a cluster around 1q21.3. The S100 family bind to p53 and regulate the oligomerization of p53.[41] Copy number gain and high expression of the S100 family is reported in relation with progression and recurrence of stomach cancer,[42] lung cancer, pancreatic cancer and brain tumor.[42,43] This suggests that S100A2 is likely to be linked with CRC.

ABHD2 is located at 15q26.1 and is a member of the abhydrolase (ABHD) superfamily. Li et al reported an association between ABHD6 and cancer,[44] but the detailed function of this gene remains uncertain. ABHD12 is located at 20p11.21 and also belongs to the abhydrolase superfamily.

OIT3 is located at 10q22.1 and is known to be related to hepatocyte formation and liver function,[45] but the details of this relationship are also not certain.

From these observations, we could extract strong candidate genes by comparing the relation between Gene Set I and Gene Set II. The number of genes in Gene Set III (14 genes) was much smaller than the number in Gene Set I (1876 genes) or Gene Set II (558 genes). When we used a cut-off value of only $P < 0.05$ for the expression analysis, many genes were extracted, but most of them seemed to be unrelated to either CNA or expression change. For this reason, we used FC as well as *P*-value for the analysis.

As described previously, not all genes in Gene Set I had a change in expression level. This indicates that the copy number itself is not parallel with the expression level. Transcription regulation is also related to the expression level, so changes in the upstream genes, promoter sequences and feedback regulators will all result in changes in the expression level. We also have to keep in mind that our observations were made using microarray analysis, so they might also be affected by alternative splicing, due to the design of the probes. Detailed research must be performed with other methods, such as Reverse Transcription Polymerase Chain Reaction (RT-PCR) or immunostaining. CNA also has to be confirmed with qPCR or FISH analysis. In addition, it is necessary to confirm the relation of the genes in Table 3 using independent CRC cases as cancer markers, and to identify the function of each gene using molecular biological experiments before these genes can be used as anti-cancer drug targets.

Pollack et al found that 62% of highly amplified genes in breast cancer exhibit at least twofold increased expression.[46] In our research, we could not observe this kind of correlation, and amplification and increased expression appeared to be rather independent. However, this was rather good for us, as it became a strong selection criterion. Stranger et al found some correlation between SNPs and CNA with the expression levels of the gene in Hapmap samples; however, the overlap of the expression change between SNPs and CNV was small.[47] This data suggests that part of the expression change is due to CNV but most of them are related to SNPs. That means the relationship between CNA and expression is rather weak.

We have analyzed not only CRC but also liver cancer and oral cancer, and observing quite different tendencies in each cancer. Among them, CRC was the most difficult cancer to analyze by single omics data using this method. We assume that breast cancer in the previous study[47] show a different relationship between copy number and expression than that of CRC.

Here, we want to emphasize that we are suggesting use of the multi-omics analysis for selecting cancer-related genes and decreasing the false positive signals that tend to come out of comprehensive analysis. CNA and gene expression have some relationship, but are not linked directly. This means we can observe the results from a different perspective. As suggested above, it is important to validate the CNA and/or expression change, but we usually obtain similar results by other methods. Also as this multidimensional analysis is a kind of validation and can reduce the false positive candidates, we think that the results can be used a primary indicator tool without validation of each gene.

We observed differences between groups T_A and T_B, especially in Gene Set III, where we classified the samples according to the invasiveness of the cancer (T1 and T2 vs. T3 and T4, and T1, T2 and T3 vs. T4). This means that the progression of the CRC resulted in changes in the genes and their expression levels.

Conclusion

We analyzed the genes related to CRC that showed CNA by comparing CNA with clinico-pathological classification. We also analyzed the genes that showed expression level changes related with CRC by a similar method. The combination of these two methods was an efficient method of selecting possible candidate genes. Each method may contain high background noise in an omics study, but when used together, especially when they are not directly related, we had a very effective result.

This method can enhance the efficiency of cancer omics analysis, and could find a new marker or target for CRC.

Acknowledgments

This work was supported by Special Coordination Funds for Promoting Science & Technology (Japan Science and Technology Agency), and a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan. We would like to thank Dr. Dana Ichinotsubo for reading and providing comments on this manuscript.

Abbreviations

5-FU: 5-fluorouracil; ARHGDIB: Rho GDP dissociation inhibitor (GDI) beta; CNA: Copy Number Aberration; CRC: Colorectal Cancer; CXCL3 chemokine (C-X-C motif) ligand 3;CXCL6: chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2);

DCC: deleted in colorectal carcinoma; FC: Fold Change; SNPs: Single Nucleotide Polymorphisms; OIT3 oncoprotein induced transcript 3; RALBP1: ralA binding protein 1; RNMT RNA (guanine-7-) methyltransferase; S100A2 S100 calcium binding protein A2; SULT1B1: sulfotransferase family, cytosolic, 1B, member 1; TYMS thymidylate synthetase; UGT2B28: UDP glucuronosyltransferase 2 family polypeptide B28.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Outline of Health, Labor and Welfare Statistics. (Ed. Ministry of Health Labor and Welfare). http://www.mhlw.go.jp/english/database/index.html. Accessed May 28, 2010.
2. Mizushima H, Mogushi K, Ohashi W, et al. TMDU clinical omics database project—integration of OMICS data and clinical information. *ISMB*. 2007:I79.
3. Tanaka H, Arii S, Sugihara K, Miki Y, Inazawa J, Mizushima H. TMDU clinical omics database—integrating OMICS data and clinical information. In: Proceedings of the *Japanese Cancer Assoc Conference*. Oct 3–5, Nagoya Japan 2007. p. 403.
4. Cancer Information Physician Data Query (PDQ®): National Cancer Institute, NIH: http://mext-cancerinfo.tri-kobe.org/database/pdq/index.html. Accessed May 28, 2010.
5. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. *Nature*. 1998 Dec 17;396(6712):643–9.
6. Kinzler KW, Vogelstein B. Landscaping the cancer terrain. *Science*. 1998 May 15;280(5366):1036–7.
7. Lindblom A. Different mechanisms in the tumorigenesis of proximal and distal colon cancers. *Curr Opin Oncol*. 2001 Jan;13(1):63–9.
8. Nakao M, Kawauchi S, Furuya T, et al. Identification of DNA copy number aberrations associated with metastases of colorectal cancer using array CGH profiles. *Cancer Genet Cytogenet*. 2009 Jan 15;188(2):70–6.
9. Jones AM, Thirlwell C, Howarth KM, et al. Analysis of copy number changes suggests chromosomal instability in a minority of large colorectal adenomas. *J Pathol*. 2007 Nov;213(3):249–56.
10. Bartos JD, Gaile DP, McQuaid DE, et al. aCGH local copy number aberrations associated with overall copy number genomic instability in colorectal cancer: coordinate involvement of the regions including BCR and ABL. *Mutat Res*. 2007 Feb 3;615(1–2):1–11. Epub 2007 Jan 2.
11. Lassmann S, Weis R, Makowiec F, et al. Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas. *J Mol Med*. 2007 Mar;85(3):293–304. Epub 2006 Dec 2.
12. Jones AM, Douglas EJ, Halford SE, et al. Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene*. 2005 Jan 6;24(1):118–29.
13. Lips EH, de Graaf EJ, Tollenaar RA, et al. Single nucleotide polymorphism array analysis of chromosomal instability patterns discriminates rectal adenomas from carcinomas. *J Pathol*. 2007 Jul;212(3):269–77.
14. Andersen CL, Wiuf C, Kruhøffer M, Korsgaard M, Laurberg S, Ørntoft TF. Frequent occurrence of uniparental disomy in colorectal cancer. *Carcinogenesis*. 2007 Jan;28(1):38–48. Epub 2006 Jun 13.
15. Alcock HE, Stephenson TJ, Royds JA, Hammond DW. Analysis of colorectal tumor progression by micro dissection and comparative genomic hybridization. *Genes Chromosomes Cancer*. 2003 Aug;37(4):369–80.
16. Aragane H, Sakakura C, Nakanishi M, et al. Chromosomal aberrations in colorectal cancers and liver metastases analyzed by comparative genomic hybridization. *Int J Cancer*. 2001 Dec 1;94(5):623–9.
17. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006 Nov 23;444(7118):444–54.
18. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: new insights in genome diversity. *Genome Res*. 2006 Aug;16(8):949–61. Epub 2006 Jun 29.
19. Midorikawa Y, Yamamoto S, Ishikawa S, et al. Molecular karyotyping of human hepatocellular carcinoma using single-nucleotide polymorphism arrays. *Oncogene*. 2006 Sep 7;25(40):5581–90. Epub 2006 Jun 19.
20. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006 Feb;7(2):85–97.
21. Cano A, Pérez-Moreno MA, Rodrigo I, et al. The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat Cell Biol*. 2000 Feb;2(2):76–83.
22. Stemmer V, de Craene B, Berx G, Behrens J. Snail promotes Wnt target gene expression and interacts with beta-catenin. *Oncogene*. 2008 Aug 28; 27(37):5075–80. Epub 2008 May 12.
23. Furne C, Rama N, Corset V, Chédotal A, Mehlen P. Netrin-1 is a survival factor during commissural neuron navigation. *Proc Nat Acad Sci U S A*. 2008 Sep 23;105(38):14465–70. Epub 2008 Sep 16.
24. Nagar S, Remmel RP. Uridine diphosphoglucuronosyltransferase pharmacogenetics and cancer. *Oncogene*. 2006 Mar 13;25(11):1659–72.
25. Lévesque E, Turgeon D, Carrier JS, Montminy V, Beaulieu M, Bélanger A. Isolation and characterization of the UGT2B28 cDNA encoding a novel human steroid conjugating UDP-glucuronosyltransferase. *Biochemistry*. 2001 Apr 3;40(13):3869–81.
26. van Coillie E, van Aelst I, Wuyts A, et al. Tumor angiogenesis induced by granulocyte chemotactic protein-2 as a countercurrent principle. *Am J Pathol*. 2001 Oct;159(4):1405–14.
27. Gijsbers K, Gouwy M, Struyf S, et al. GCP-2/CXCL6 synergizes with other endothelial cell-derived chemokines in neutrophil mobilization and is associated with angiogenesis in gastrointestinal tumors. *Exp Cell Res*. 2005 Feb 15;303(2):331–42.
28. Zhu YM, Bagstaff SM, Woll PJ. Production and upregulation of granulocyte chemotactic protein-2/CXCL6 by IL-1beta and hypoxia in small cell lung cancer. *Br J Cancer*. 2006 Jun 19;94(12):1936–41. Epub 2006 May 23.
29. Rubie C, Frick VO, Wagner M, et al. ELR+ CXC chemokine expression in benign and malignant colorectal conditions. *BMC Cancer*. 2008 Jun 25;8:178.
30. Enokizono J, Kusuhara H, Sugiyama Y. Regional expression and activity of breast cancer resistance protein (Bcrp/Abcg2) in mouse intestine: overlapping distribution with sulfotransferases. *Drug Metab Dispos*. 2007 Jun;35(6):922–8. Epub 2007 Mar 12.
31. Milano F, Jorritsma T, Rygiel AM, et al. Expression pattern of immune suppressive cytokines and growth factors in oesophageal adenocarcinoma reveal a tumour immune escape-promoting microenvironment. *Scand J Immunol*. 2008 Dec;68(6):616–23.
32. Bièche I, Chavey C, Andrieu C, et al. CXC chemokines located in the 4q21 region are up-regulated in breast cancer. *Endocr Relat Cancer*. 2007 Dec;14(4):1039–52.
33. Li A, Varney ML, Singh RK. Constitutive expression of growth regulated oncogene (GRO) in human colon carcinoma cells with different metastatic potential and its role in regulating their metastatic phenotype. *Clin Exp Metastasis*. 2004;21(7):571–9.
34. Singhal SS, Singhal J, Yadav S, et al. Regression of lung and colon cancer xenografts by depleting or inhibiting RLIP76 (Ral-binding protein 1). *Cancer Res*. 2007 May 1;67(9):4382–9.

35. Sharma R, Hoskins JM, Rivory LP, et al. Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphisms and toxicity to capecitabine in advanced colorectal cancer patients. *Clin Cancer Res*. 2008 Feb 1; 14(3):817–25.
36. Mosesson Y, Mills GB, Yarden Y. Derailed endocytosis: an emerging feature of cancer. *Nat Rev Cancer*. 2008 Nov;8(11):835–50.
37. Iida H, Noda M, Kaneko T, Doiguchi M, Mōri T. Identification of RAB12 as a vesicle-associated small GTPase highly expressed in Sertoli cells of rat testis. *Mol Reprod Dev*. 2005 Jun;71(2):178–85.
38. McCracken S, Fong N, Rosonina E, et al. 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev*. 1997 Dec 15;11(24):3306–18.
39. Schunke D, Span P, Ronneburg H, et al. Cyclooxygenase-2 is a target gene of rho GDP dissociation inhibitor beta in breast cancer cells. *Cancer Res*. 2007 Nov 15;67(22):10694–702.
40. Ota T, Maeda M, Suto S, Tatsuka M. LyGDI functions in cancer metastasis by anchoring Rho proteins to the cell membrane. *Mol Carcinog*. 2004 Apr;39(4):206–20.
41. Fernandez-Fernandez MR, Rutherford TJ, Fersht AR. Members of the S100 family bind p53 in two distinct ways. *Protein Sci*. 2008 Oct;17(10):1663–70. Epub 2008 Aug 11. PMID: 18694925.
42. Liu J, Li X, Dong GL, et al. In silico analysis and verification of S100 gene expression in gastric cancer. *BMC Cancer*. 2008 Sep 16;8:261. PMID: 18793447.
43. Rand V, Prebble E, Ridley L, et al. Children's Cancer and Leukaemia Group Biological Studies Committee. Investigation of chromosome 1q reveals differential expression of members of the S100 family in clinical subgroups of intracranial paediatric ependymoma. *Br J Cancer*. 2008 Oct 7; 99(7):1136–43. Epub 2008 Sep 9.
44. Li F, Fei X, Xu J, Ji C. An unannotated alpha/beta hydrolase superfamily member, ABHD6 differentially expressed among cancer cell lines. *Mol Biol Rep*. 2008 Mar 22. [Epub ahead of print].
45. Xu ZG, Du JJ, Zhang X, et al. A novel liver-specific zona pellucida domain containing protein that is expressed rarely in hepatocellular carcinoma. *Hepatology*. 2003 Sep;38(3):735–44.
46. Pollack JR, Sørlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Nat Acad Sci U S A*. 2002 Oct 1; 99(20):12963–8.
47. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007 Feb 9;315(5813):848–53.