GENOME LETTERS 2002, 1, 26–34
www.aspbs.com/genomelett

RESEARCH ARTICLE

Gene Expression Profiling of Metastatic and Nonmetastatic Colorectal Cancer Cell Lines

Matthias Futschik,[1] Aaron Jeffs,[2] Sharon Pattison,[3] Nikola Kasabov,[1]
Michael Sullivan,[3] Arend Merrie,[4] and Anthony Reeve[2,3,*]

[1] *Knowledge Engineering Laboratory, Department of Information Science, University of Otago, Dunedin, New Zealand*
[2] *Otago Genomics Facility, Department of Biochemistry, University of Otago, Dunedin, New Zealand*
[3] *Cancer Genetics Laboratory, Department of Biochemistry, University of Otago, Dunedin, New Zealand*
[4] *Department of Surgery, Auckland Hospital, New Zealand*

(Received: 27 December 2001; revised/accepted: 29 January 2002)

ABSTRACT: cDNA microarrays were used to compare the gene expression pattern of a nonmetastatic colorectal cancer cell line (SW480) with its metastatic derivative (SW620). Co-hybridization of fluorescently labeled cDNA generated from SW480 (Cy3) and SW620 (Cy5) total RNA to microarrays containing 4000 human cDNA clones revealed differential expression of 129 genes involved in the regulation of transcription, cell-cycle control and division, cell signaling, cell adhesion, and cell metabolism. The results of this microarray analysis corresponded to previously reported gene expression profiling experiments with SW480 and SW620 using SAGE. Predictably, the metastatic cell line SW620 exhibited underexpression of genes involved in cell adhesion and overexpression of genes involved in transcription and translation compared with its nonmetastatic counterpart SW480. Finally, by applying a novel bioinformatics approach to the gene expression data, a number of genes underexpressed in SW620 compared with SW480 were demonstrated to map to chromosomal band 17q21-23, a region that may be associated with loss of heterozygosity during the metastatic progression of colorectal cancer.

Keywords: Colorectal Cancer, Metastasis, Gene Expression, Profiling, Microarrays, Bioinformatics.

1. INTRODUCTION

Colorectal cancer (CRC) represents a significant global health problem. At least 50% of people in the Western world will develop a colorectal adenoma by the age of 50, with perhaps one in 10 progressing to malignancy [1]. There are large geographic differences in the incidence of CRC, with high rates in the United States, Australia, and New Zealand, and low rates in Japan, South America, and Africa [2]. New Zealand and Australia have the highest rates of CRC in the world for both sexes [3], and non-Maori New Zealand women have the highest age-standardized rate of CRC in the world [3].

If detected early, nonmetastatic CRC can be surgically resected with a favorable prognosis. However, CRC metastases are associated with a poor prognosis and are often refractory to chemotherapy. Because CRC has defined precursor lesions and exhibits slow progression to metastasis, it should be an essentially preventable disease with appropriate screening procedures. However, current screening tools for CRC in asymptomatic people lack appropriate sensitivity and specificity and are not universally accepted for large-scale use in the general population. A molecular genetic approach may identify people at high risk for CRC and may allow for targeted, less invasive, more cost-effective screening [4]. Furthermore, the current clinicopathological classification scheme does not necessarily provide an accurate indication of response to treatment and survival. The underlying genetic events that lead to the progression of CRC from adenomatous polyp to malignant tumor is described by the adenoma–carcinoma sequence.

*Author to whom correspondence should be addressed.

1537-3053/2002/01/26/09/$17.00 +.25 doi:10.1166/gl.2002.005

Although much is known about the adenoma–carcinoma sequence, very little is known about the transition of CRC from nonmetastatic to metastatic disease. Here we present data from cDNA microarray assays that compare gene expression profiles between primary and metastatic CRC cell lines derived from the same patient [5]. Genes with a different expression pattern between metastatic and nonmetastatic disease may have an important role in the progression from localized to disseminated cancer and may therefore be of clinical use as prognostic markers for colorectal cancer.

2. MATERIALS AND METHODS

2.1. Cell Lines and RNA Extraction

SW480 and SW620 colorectal cancer cell lines were obtained from the American Type Culture Collection and grown in alpha minimum essential medium containing 10% fetal bovine serum. RNA was extracted from the cells with Trizol (Life Technologies) and further purified with the use of RNeasy columns (Qiagen) according to the manufacturer's instructions.

2.2. cDNA Microarrays

Four thousand sequence-verified human cDNA clones (Research Genetics, release GF211) were printed in duplicate onto poly-L-lysine-coated slides with the use of an SDDC-2 arrayer (ESI, Toronto) equipped with Telechem SMP3 quill pins and post-processed as described [6]. Fluorescently labeled cDNA was generated from 20 μg total RNA by the amino-allyl labelling method as described (www.microarrays.org) and hybridized to the arrays in UltraHyb buffer (Ambion) at 50 °C with a GeneTAC Hybridization Station (Genomic Solutions). After 20 h of hybridization the slides were washed as described [6], dried by centrifugation at 600 rpm for 6 min, and then scanned with a ScanArray 5000 Microarray Analysis System (GSI Lumonics). Fluorescent intensity data were extracted with QuantArray microarray analysis software (version 2.1, GSI Lumonics). Four replicate experiments were performed, with separate labeling reactions hybridized to different microarrays. Control spots were printed with foreign DNA (COT-1, Rice, SS-DNA, Lambda, Apple) as well as negative control spots (3× SSC buffer only) to assess the background threshold of fluorescence.

2.3. Data Pre-processing and Analysis

Genes showing low fluorescence were excluded from further data analysis, as plots of Cy3 versus Cy5 signal intensity indicated a strong bias toward the Cy3 channel for genes with low intensities (data not shown). This nonlinear behavior of the signals would interfere with the global normalization procedure subsequently used. Therefore, a threshold was derived below which the data were considered unreliable by analysis of the intensities of negative control spots (3× SSC buffer only). Although the fluorescence of these spots were locally background corrected by the scanning software, negative control spots still showed low fluorescence intensities. Because these intensities are based on nonspecific hybridization effects, we considered genes with similar intensities as not expressed. To achieve a robust estimate, the logged intensity values of the negative control spots were iteratively fitted by a normal distribution. Deriving the mean intensity value μ_{nc} and the standard deviation s_{nc}, we used $\mu_{nc} + 2s_{nc}$ as the *expression* threshold. A gene was excluded from further analysis if 50% of the spots representing the gene had intensities below the expression threshold in both channels across all slides. The genes remaining for the subsequent analysis numbered 3724. The expression values for these genes were globally normalized to generate equal total intensities in the two channels. To balance the data, the intensity ratios were log-transformed, providing a symmetrical distribution of the ratios around zero (Fig. 1).

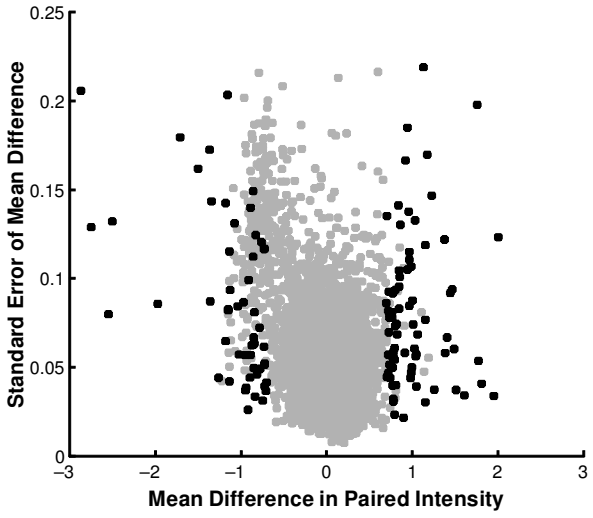

Figure 1. Distribution of mean differences of paired intensities $\overline{\Delta}$ vs. standard error for the mean differences $s_{\overline{\Delta}}$ for the complete set of genes in our study. Genes that showed intensities above the expression threshold and scored a p value of less than 0.01 are displayed as black dots. Genes that are below the expression threshold or have insignificant fold changes are displayed as gray spots. Note that the mean of differences $\overline{\Delta}$ is shown on a logarithmic scale with base 2 in the figure. Genes with $\overline{\Delta} = 1$ therefore have an average fold change of 2.

2.4. Statistical Evaluation of Differential Gene Expression

The differential expression of genes was selected as statistically significant with a paired t-test. This t-test compares the average fold change with the error of the average fold change and assigns a p value to this change based on Student's t distribution. A large average fold change combined

with a small error of the fold change yields a high significance. This paired t-test has the form

$$t = \frac{\overline{\Delta}}{s_{\overline{\Delta}}}$$

where $\overline{\Delta}$ is the mean of the logged fold changes and $s_{\overline{\Delta}}$ is the standard error of the mean of the logged fold changes. The standard error is defined by

$$s_{\overline{\Delta}} = \sqrt{\sum_i (d_i - \bar{d})/n(n-1)}$$

where d_i are the logged fold changes, $\bar{d}$ is the mean of the logged fold changes, and n is the number of replicates. We observed in our study, however, that genes of low intensity and small fold changes frequently displayed a high significance based on this original t-test because of a very low standard error. This is to be expected since the standard error is itself a statistic variable with a broad distribution, especially for a low number of replicates. Any undetected experimental bias in the expression values can strongly interfere with the paired t-test for significance, particularly for low expression values.

To obtain a more robust estimate of the error, we therefore modified its derivation. In our model, the total error for each gene $\sigma_{\text{tot,gene}}$ consists of the specific variability of a gene σ_{gene} and a general background noise in the experiment $\sigma_{\text{exp,}}$ so that $\sigma^2_{\text{tot,gene}} = \sigma^2_{\text{gene}} + \sigma^2_{\text{exp}}$. We estimated the general experimental error σ_{exp} by calculating the median variance of the logged ratios for genes below the expression threshold. This provides a measure for random fluctuations of the intensity values in our experiment. Replacing $s_{\overline{\Delta}}$ with σ_{tot} in the paired t-test formula leads to a robust test for significance, as our analysis showed. The error $\sigma_{\text{tot,gene}}$ essentially sets an estimated minimal possible error in the experiment, while it also includes the gene-specific variations of expression and hybridization. A similar modification of the t-test was recently proposed by Tusher and Tibshirani [7].

2.5. Correlation of Differential Gene Expression with Chromosomal Loci

To determine the significance of finding a relatively large number of differentially expressed genes within a small chromosomal location, we calculated the probability that a specific number of significantly differentially expressed genes were located by chance within a cytoband, using the following formula. The p values were calculated with the use of the hypergeometric distribution

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{s}{i}\binom{g-s}{n-i}}{\binom{g}{n}}$$

where g is the total number of genes in the experiment, n is the total estimated number of genes within the cytoband, and s is the total number of differentially expressed genes. The hypergeometric distribution gives the probability that through n random drawings from a pool of g genes at least k genes will be drawn out of a specific subset of s genes. This approach was previously used to assess the statistical significance of the enrichment of functional categories gene expression clusters within [8] but to date has not been applied to map gene expression to chromosomal domains.

3. RESULTS

3.1. Differential Gene Expression in SW620 Compared with SW480

Gene expression profiles for the SW480 and SW620 cell lines were generated as described in the Experimental Details. The microarray data were filtered to exclude genes below the expression threshold before statistical tests for significance were applied. In the first instance we used an unmodified paired t-test. The results of this significance test were judged to be unreliable, since a large number of genes of low intensity and small fold changes were detected as significant because of their very small errors. A comparison of these errors showed that they underestimated the general experimental error in our experiment, as defined in the Experimental Details (data not shown). For a robust assessment of the experimental error we modified the error estimation for the t-test. Using this modified significance test, we found 129 genes with a significance level of less than 0.01. Seventy-five genes were significantly overexpressed, and 54 genes were significantly underexpressed ($p < 0.01$, Tables I and II, respectively). These results are shown in Figure 1. The mean fold changes versus the error of the mean fold changes are displayed. Significant differentially expressed genes show large fold changes compared with the errors. The distribution indicates an asymmetry for genes with a large error caused by the bias toward Cy3 for low intensities. Genes in this intensity regime were excluded from our significance test by application of the expression threshold (see Experimental Details).

Figure 1 further shows that the underexpressed and overexpressed genes in SW620 compared with SW480 are symmetrically located and well distinguished from the rest of the genes that fell below the threshold of significance.

3.2. Differential Gene Expression and Correlation with Chromosomal Location

By examining the gene expression data in Table II, there was an apparent nonrandom chromosomal clustering of genes that were underexpressed in the SW620 cell line. To determine whether this association was nonrandom, a statistical approach was applied.

The set of cDNA clones used in this study contained 52 genes located in the cytoband s17q21-22 and 17q21-23.

Table 1. Genes up-regulated in the SW620 cell line.

GenBank no.	Gene name	Gene product	Average fold change	Cytoband	*p* value
H19203	PRDX3	Peroxiredoxin 3	3.9	10q25-q26	$2.4 10^{-5}$
N74131	TFF3	Trefoil factor 3 (intestinal)	3.8	21q22.3	$4.1 10^{-5}$
AA487893	TM4SF1	Transmembrane 4 superfamily member 1	3.5	3q21-q25	$2.0 10^{-5}$
AA857343	TAF2N	TATA box binding protein (TBP)-associated factor, RNA polymerase II, N, 68 kDa (RNA-binding protein 56)	3.4	17q11.1-q11.2	$4.6 10^{-5}$
AA873060	LAP18	Leukemia-associated phosphoprotein p18 (stathmin)	3.0	1p36.1-p35	$8.7 10^{-5}$
AA011215	SAT	Spermidine/spermine N1-acetyltransferase	2.8	Xp22.1	0.00013
AA453420		ESTs	2.8		0.00014
R28294	GCSH	Glycine cleavage system protein H	2.7	16	0.00015
AA495936	MGST1	Microsomal glutathione *S*-transferase 1	2.6	12p12.3-p12.1	0.00020
AA634008	RPS23	Ribosomal protein S23	2.6	5q	0.00017
AA779165	ARL4	ADP-ribosylation factor-like 4	2.6	7p21-p15.3	0.00023
AA486220	KARS	Lysyl-tRNA synthetase	2.5	16q23-q24	0.000236
AA135152	GPX2	Glutathione peroxidase 2 (gastrointestinal)	2.4	14q24.1	0.00074
AA099534	PC4	Activated RNA polymerase II transcription cofactor 4	2.2	8	0.00067
AA485353	LGALS3BP	Galectin 6 binding protein	2.2	17q25	0.00068
AA862465	AZGP1	Alpha-2-glycoprotein 1, zinc	2.2	7q22.1	0.00059
AA644234	ATP5C1	ATP synthase, mitochondrial F1 complex, gamma polypeptide 1	2.1	10q22-q23	0.00110
H79534	HBE1	Hemoglobin, epsilon 1	2.1	11p15.5	0.00067
AA055350	ADORA2B	Adenosine A2b receptor	2.1	17p12-p11.2	0.00122
AA868008	H4FG	H4 histone family, member G	2.1	6p21.3	0.00175
T71363	SDS	Serine dehydratase	2	12	0.00126
R93124	AKR1C1	Aldo-keto reductase family 1, member C1	2	10p15-p14	0.00177
AA598974	CDC2	Cell division cycle 2, G_1 to S and G_2 to M	2	10q21.1	0.00075
N70463	BTG1	B-cell translocation gene 1	2	12q22	0.00119
AA156571	AARS	Alanyl-tRNA synthetase	2	16q22	0.00160
AA625981	FKBP1A	FK506-binding protein 1A (12kD)	2	20p13	0.00137
AA450265	PCNA	Proliferating cell nuclear antigen	2	20pter-p12	0.00147
H59916	CD24	CD24 antigen	2	6q21	0.00169
AA669674	EIF3S6	Eukaryotic translation initiation factor 3, subunit 6 (48 kDa)	2	8q22-q23	0.00154
R83000	BTF3	Basic transcription factor 3	1.9	5	0.00187
AA485214	NUCB2	Nucleobindin 2	1.9	11p15.1-p14	0.00277
AA497029	LDHA	Lactate dehydrogenase A	1.9	11p15.4	0.00259
AA478273	APEX	APEX nuclease (multifunctional DNA repair enzyme)	1.9	14q11.2-q12	0.00186
H22922	MFNG	Manic fringe (*Drosophila*) homolog	1.9	22q12	0.00824
AA609284	EPHB6	EphB6	1.9	7q33-q35	0.00220
AA670134	RPL8	Ribosomal protein L8	1.9	8q	0.00164
AA464600	MYC	v-myc avian myelocytomatosis viral oncogene homolog	1.9	8q24.12-q24.13	0.00190
AA477400	TPM2	Tropomyosin 2 (beta)	1.9	9p13.2-p13.1	0.00128
AA446820	OAT	Ornithine aminotransferase	1.8	10q26	0.00390
AA469965	LCK	Lymphocyte-specifi c protein tyrosine kinase	1.8	1p35-p34.3	0.00196
W02101	HNRPA2B1	Heterogeneous nuclear ribonucleoprotein A2/B1	1.8	7p15	0.00367
N33274	ADE2H1	Polypeptide similar to SAICAR synthetase and AIR carboxylase	1.7	8	0.00553
AA455126	ATP5G2	ATP synthase, mitochondrial F0 complex, subunit c, isoform 2	1.7	12	0.00553
AA679177	HSPC121	Butyrate-induced transcript 1	1.7	15	0.00453
AA706968	ZWINT	ZW10 interactor	1.7	10q21-q22	0.00385
AA459909	PCBD	6-Pyruvoyl-tetrahydropterin synthase	1.7	10q22	0.00634
AA629641	RPS13	Ribosomal protein S13	1.7	11p	0.00474
AA464147	CARS	Cysteinyl-tRNA synthetase	1.7	11p15.5	0.00605
AA126911	HNRPA1	Heterogeneous nuclear ribonucleoprotein A1	1.7	12q13.1	0.00507
AA676998	TMPO	Thymopoietin	1.7	12q22	0.00612
AA629808	RPL6	Ribosomal protein L6	1.7	12q23-24.1	0.00660
W73144	LCP1	Lymphocyte cytosolic protein 1 (L-plastin)	1.7	13q14.3	0.00575
AA700604	SORD	Sorbitol dehydrogenase	1.7	15q15.3	0.00614
AA872341	RPS15A	Ribosomal protein S15a	1.7	16p	0.00387
AA682613	CFDP1	Craniofacial development protein 1	1.7	16q22.2-q22.3	0.00876
AA486072	SCYA5	Small inducible cytokine A5 (RANTES)	1.7	17q11.2-q12	0.00517
AA680244	RPL11	Ribosomal protein L11	1.7	1p36.1-p35	0.00658
N75581	FUBP1	Far upstream element (FUSE) binding protein 1	1.7	1pter-p22.2	0.00546
AA480995	MTHFD2	Methylene tetrahydrofolate dehydrogenase (NAD + dependent)	1.7	2p24.3-p24.1	0.00599
AA450205	TLOC1	Translocation protein 1	1.7	3q26.2-q27	0.00356
AA629987	PPID	Peptidylprolyl isomerase D (cyclophilin D)	1.7	4q31.3	0.00607

Continued

Table 1. Continued

GenBank No.	Gene name	Gene Product	Average fold change	Cytoband	*p* value
W96058	HNRPH1	Heterogeneous nuclear ribonucleoprotein H1 (H)	1.7	5q35.3	0.00775
AA410517	SERPINB6	Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 6	1.7	6p25	0.00609
H73724	CDK6	Cyclin-dependent kinase 6	1.7	7q21-q22	0.00503
AA490477		*Homo sapiens*, clone IMAGE:3357927, mRNA, partial cds	1.7		0.00623
AA486761	YARS	Tyrosyl-tRNA synthetase	1.6	1	0.00876
AA598836	CUL4A	Cullin 4A	1.6	13	0.00873
AA133212	NCOA4	Nuclear receptor coactivator 4	1.6	10q11.2	0.00852
AA032090	DDX11	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 11	1.6	12p11	0.00797
AA669452	EIF2S1	Eukaryotic translation initiation factor 2, subunit 1 (alpha, 35kD)	1.6	14q21.1-q22.3	0.00859
N70794	ACADM	Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	1.6	1p31	0.00849
AA463497	MCP	CD46, trophoblast-lymphocyte cross-reactive antigen	1.6	1q32	0.00810
AA701455	CENPF	Centromere protein F (350/400 kDa, mitosin)	1.6	1q32-q41	0.00968
H84048	RBL1	Retinoblastoma-like 1 (p107)	1.6	20q11.2	0.00996

Table 2. Genes down-regulated in the SW620 cell line.

GenBank no.	Gene name	Gene product	Average fold change	Cytoband	*p* value
R55303	NGFR	Nerve growth factor receptor	8.2	17q21-q22	2.3410^{-5}
H25546	SAA1	Serum amyloid A1	7.1	11p15.1	2.5410^{-5}
AA487486	CCND1	Cyclin D1	6	11q13	4.2310^{-5}
W60057	KRT13	Keratin 13	6	17q21-q23	4.7310^{-5}
R33154		ESTs	4		2.36E−05
AA160507	KRT5	Keratin 5	3.6	12q13	0.000271
AA873604	CRIP1	Cysteine-rich protein 1 (intestinal)	3.1	7q11.23	0.000318
AA428778	EFNB1	Ephrin-B1	2.9	Xq12	0.000812
T53298	IGFBP7	Insulin-like growth factor binding protein 7	2.8	4q12	0.00029
AA464731	S100A11	S100 calcium-binding protein A11 (calgizzarin)	2.4	1q21	0.000488
AA865464	LY6E	Lymphocyte antigen 6 complex, locus E	2.4	8q24.3	0.000396
AA424695	ITGA3	Integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	2.3	17	0.00075
H59614	IGF2	Insulin-like growth factor 2	2.3	11p15.5	0.000695
AA629265	NR1H2	Nuclear receptor subfamily 1, group H, member 2	2.3	19q13.3-19q13.3	0.000707
AA496997	LMNA	Lamin A/C	2.3	1q21.2-q21.3	0.000764
AA485362	GPX1	Glutathione peroxidase 1	2.3	3p21.3	0.000595
AA458878		*Homo sapiens*, clone IMAGE:3544662, mRNA, partial cds	2.3		0.006899
H15456	CAPN1	Calpain 1, (mu/I) large subunit	2.2	11q13	0.001085
AA610066	HOXB6	Homeo box B6	2.2	17q21-q22	0.000746
AA418036	GLI3	GLI3, GLI-Kruppel family member	2.2	7p13	0.00104
H51645	ACTN4	Actinin, alpha 4	2.1	19q13	0.001366
AA460393		*Homo sapiens* pinch-2 protein	2.1		0.001252
T46897	GP110	Cell membrane glycoprotein, 110000M(r) (surface antigen)	2	20pter-20p12.1	0.001852
AA486239	FLNB	Filamin B, beta (actin-binding protein-278)	2	3p14.3	0.002093
N68166	GNB2	Guanine nucleotide binding protein (G protein), beta polypeptide 2	2	7q22	0.004786
AA485052		ESTs	2		0.003006
AA464729	E2-EPF	Ubiquitin carrier protein	1.9	17p12-p11	0.003552
AA857101	HOXB7	Homeo box B7	1.9	17q21-q22	0.002071
AA479882	KRT10	Keratin 10	1.9	17q21-q23	0.002181
AA421518	AP2S1	Adaptor-related protein complex 2, sigma 1 subunit	1.9	19q13.2-q13.3	0.002515
W73810	EMP3	Epithelial membrane protein 3	1.9	19q13.3	0.003025
N54551	PPP5C	Protein phosphatase 5, catalytic subunit	1.9	19q13.3	0.004316
H79047	IGFBP2	Insulin-like growth factor binding protein 2	1.9	2q33-q34	0.002681
AA418694	ATOX1	ATX1 (antioxidant protein 1, yeast) homolog 1	1.9	5q32	0.002847
AA598776	CDC20	CDC20	1.9	9q13-q21	0.003262
AA478724	IGFBP6	Insulin-like growth factor binding protein 6	1.8	12q13	0.003705
AA026118	GRP	Gastrin-releasing peptide	1.8	18q21.1-q21.32	0.003916
AA599177	CST3	Cystatin C	1.8	20p11.2	0.004067
AA700832	RBP1	Retinol-binding protein 1	1.8	3q23	0.006578
N53485	ADD1	Adducin 1 (alpha)	1.8	4p16.3	0.003547
H70775	H2BFL	H2B histone family, member L	1.8	6p21.3	0.003803

Continued

Table 2. Continued

GenBank no.	Gene name	Gene product	Average fold change	Cytoband	*p* value
AA489314	HSGP25L2G	gp25L2 protein	1.8		0.005646
AA464417	IFITM3	Interferon-induced transmembrane protein 3 (1-8U)	1.7	8	0.007379
AA167728	SIVA	CD27-binding (Siva) protein	1.7	22	0.008229
N55480	HRMT1L2	HMT1 (hnRNP methyltransferase, *S. cerevisiae*)-like 2	1.7	19q13.3	0.00856
N34827	GUSB	Glucuronidase, beta	1.7	7q21.11	0.006514
AA456051	UGCG	UDP-glucose ceramide glucosyltransferase	1.7	9q31	0.005906
AA625628	DXS1357E	Accessory proteins BAP31/BAP29	1.7	Xq28	0.008419
AA459363	HSRNASEB	seb4D	1.7		0.008458
AA888148	TUBB2	Tubulin, beta, 2	1.7		0.008431
AA630374	DUSP6	Dual specificity phosphatase 6	1.6	12q22-q23	0.009407
AA010400	ETV4	ets variant gene 4 (E1A enhancer-binding protein, E1AF)	1.6	17q21	0.008454
W73473	BMP7	Bone morphogenetic protein 7 (osteogenic protein 1)	1.6	20q13	0.009981

Six of these genes were found to be significantly underexpressed by the *t*-test analysis. A hypergeometric distribution was applied to calculate the probability that the accumulation of significantly underexpressed genes in 17q21-22 and 17q21-23 occurred merely by chance. We derived a p value of 7.67×10^{-5}, given the following parameter values: $g = 3724, n = 52, s = 53$, and $k = 6$. This extremely low p value indicates that the cluster of underexpressed genes on chromosome 17q21-23 has a very high statistical significance.

The arrayed clone set also contained 94 genes located in the cytoband 19q13.2-q13.3, of which six were significantly under-expressed. By using values of $g = 3724, n = 94$, $s = 53$, and $k = 6$, a p value of 0.0019 was calculated. This value also passes the test for a statistical significance of 0.01.

4. DISCUSSION

A metastatic cell line (SW620) and nonmetastatic cell line (SW480), both derived from the same individual, have been used as a model system for identifying genes with a potential role in colorectal metastasis. To determine the statistical significance for differentially expressed genes, a robust estimation method was used. By the derivation of an expression threshold, the error caused by the background noise in the experiment was assessed. This procedure provided an index for the quality of the microarrays, an issue that is still not resolved in the analysis of gene expression data. The estimation of the error in this analysis included a general experimental error and a gene-specific error. The derivation of this error led to a more robust assessment of the statistical significance compared with the unmodified *t*-test that assigns significance values for each gene independently.

SW480 cells have an epithelial-like morphology, in contrast to SW620 cells, which form smaller aggregates and have an elongated morphology, reflecting the initial histology of the tumor [9]. Although gene expression profiles obtained from cell lines grown *in vitro* can only provide an approximation of the biology of human tumors, a number of genes differentially up-regulated in SW620 were consistent with the biological progression of colorectal cancer. In a previous SAGE analysis of the SW480 and SW620 cell lines, three genes (keratin K5, cystatin S, and serum amyloid A) were confirmed by Northern blot to be down-regulated in SW620 compared with SW480 [10]. These same three genes were present within the 54 genes differentially down-regulated in SW620 and provide supporting validation for the microarray data presented here.

A number of genes differentially up-regulated in SW620 were consistent with the more aggressive phenotype of this cell line. Of the 75 genes differentially up-regulated in SW620, peroxiredoxin 3 (Prx3) had the highest differential expression (Table I). Prx3 has properties which may contribute to the increased metastatic potential of the SW620 cell line. For example, Prx3 regulates several signal transduction pathways that influence cell growth and apoptosis, and it is involved in protecting cells from radiation-induced cell death [11, 12]. TFF3 was similarly highly expressed. TFF3 has been shown to promote migration of intestinal cells *in vitro*, and stimulation of the HT29 and ACT116 colorectal cell lines has been shown to lead to a down-regulation of E-cadherin expression [13]. TFF3 has also been shown to protect colorectal cell lines from apoptosis, which demonstrates the potential for TFF3 expression to provide resistance to chemotherapy [13], and this led to the suggestion that TFF3 expression may result in a more aggressive phenotype [14]. The TATA-binding protein has been shown to associate with β-catenin phosphorylated at Tyr-654 [15]. Given the central role of β-catenin in signaling from the adhesion molecule E-cadherin, it is conceivable that elevated levels of TATA-binding protein may stabilize β-catenin when E-cadherin is down-regulated. Stathmin decreases microtubule stability [16], and in breast carcinomas increased expression has been associated with a higher proportion of aneuploid cells, proliferating cell nuclear antigen, and tumor size [17]. There are no reports of stathmin expression in colorectal carcinoma, and in view of its relative overexpression in the SW620 cell line, further investigation of tumors is warranted. Activated RNA polymerase II transcription cofactor 4 is a DNA binding protein which activates transcription, and increased expression in liver metastatic prostate lesions has been described

[18]. Cyclin-dependent kinase-2 phoshorylates and inactivates Rb and is overexpressed in several human malignancies. In a preliminary investigation no relationship between CDC2 and colorectal tumor grade was observed [19]. However, in breast cancer CDC2 serves as an independent predictor for the recurrence of node-negative breast cancer [20], and a multivariate analysis demonstrated that it is the second most significant prognostic variable, following lymph node status [21]. Proliferating cell nuclear antigen was originally identifed as a nuclear protein whose expression correlated with the cellular proliferative state [22]. Primary colorectal tumors strongly positive for proliferating cell nuclear antigen have a higher rate of liver metastases than cases with either no expression or a low expression [23, 24]. It is of interest that a cluster of ribosomal proteins was more highly expressed in the SW620 cell line. Ribosomal proteins have been found to be overexpressed in colorectal cancer [25], and the S13 ribosomal protein is expressed in actively growing cells [26]. In an initial trial experiment, S13 ribosomal protein mRNA was detected in a number of colorectal tumors, although its role in tumor invasion and metastasis remains to be established [26].

Several genes differentially down-regulated in SW620 have functions which may contribute to the metastatic phenotype of colorectal cancer. Of 54 genes down-regulated in SW620 compared with SW480, nerve growth factor receptor (NGFR) had the lowest expression. NGFR has apoptotic and antiproliferative effects in some types of tumor cells, although NGFR expression has not yet been examined in primary colorectal tumors. In breast cancer NGFR expression is associated with low histological grade and longer disease-free survival [27, 28]. Similarly, in prostate cancer NGFR was shown to undergo a progressive loss of expression during malignant progression, while reintroduction of NGFR into a cell line induced apoptosis [29]. Serum amyloid A1 had a 7-fold reduction in expression in SW620 compared with SW480, in aggreement with a previous SAGE analysis [10]. SAA proteins are involved in mucosal defense and inflammation associated with enteric bacteria, and although their role in colorectal cancer is unclear, higher concentrations of SAA1 have usually been associated with more advanced and metastatic disease [30–32]. Three members of the cytokeratin gene family (5, 10, and 13 were) expressed at lower levels in SW620 compared with SW 480. Cytokeratin 5 has previously been shown to have lower relative expression in SW620 [10]. Although several other members of the cytokeratin family are expressed in colorectal cancer, there are no reports analyzing the potential role cytokeratins 5, 10, and 13 may have in colorectal cancer progression. These genes are expressed however, at low levels in some malignancies. For example, in breast cancer cytokeratin 5 is produced in normal tissue, in contrast to the neoplastic component [33]. Cytokeratins 10 and 13 are expressed in normal cervical epithelia and several squamous epithelia but are virtually absent from invasive carcinomas [34, 35]. Cytokeratin 5 is also expressed in chemically transformed rat bladder cell lines, whereas it is absent from the normal cell line counterpart [36]. Three members of the insulin-like growth factor binding protein family (IGFBPs, 2, 6 and 7) were down-regulated in SW620. The IGFBPs regulate IGF availablity and modulate binding to the IGF receptors, but the exact role of the IGFBPs in colorectal cancer remains to be established. Elevated levels of IGFBP-2 have been reported to be increased with tumor agressiveness [37]. In contrast, in colon cell lines it has been suggested that IGFBP-2 might inhibit the IGF mitogenic signal [38, 39]. IGFBP7 has been found to be down-regulated at the transcription level in mammary carcinoma cell lines [40], and expression has been shown to decrease in prostate tumor epithelial cells as they progress from the benign to the malignant phenotype [41]. Cysteine-rich intestinal protein-1 (CRIP-1) was expressed at a 3-fold lower level in SW620 compared with SW480. CRIP-1 is highly expressed in the intestinal epithelium [42], and its lower relative expression in SW620 might be related to the less differentiated state of this cell line. Hashida et al. [43] demonstrated that integrin, alpha-3 had lower expression in metastatic compared with nonmetastatic colorectal tumors. The 2.3-fold lower expression of integrin, alpha-3 in SW620 might be related to the increased metastatic potential of this cell line. The expression of integrin, alpha-3 has been shown to inversely correlate with the degree of adhesion for cells grown *in vitro* [44]. However, the exact role of integrin, alpha-3 is still not clear, and other reports suggest that it is involved in increasing cell motility [45, 46].

An unexpected finding in this investigation was the clustering within two discrete chromosomal regions of a number of genes which were down-regulated in SW620. To investigate the association between chromosomal location and the clustering of differentially expressed genes, we devised a novel approach for the analysis of large-scale gene expression data. Based on a hypergeometrical distribution, chromosome cytobands 17q21-23 and 19q13.2-q13.3 showed a statistically significant overrepresentation of underexpressed genes in SW620. This finding is of interest because of the potential role that allelic loss of these regions plays in the progression of colorectal cancer. Loss of heterozygosity of the 17q21-23 region has been shown to be positively associated with lymph node metastasis [47], liver metastasis, prognosis [48], and allelic deletions of the Nm23-H1 tumor suppressor gene [49]. In contrast, other investigators have found no relationship between 17q21-23 deletions and tumor stage or metastatic behavior [50–52]. Allelic deletions of 17q21-23 were not measured in SW260 in this investigation, and although there are no reports of deletions from the cytogenetic literature, a recent investigation using spectral karyotyping described a complex chromosome 17 translocation with an unidentified breakpoint [53].

5. CONCLUSION

Chromosome cytobands 19q13.2-13.3 also had an overrepresentation of genes underexpressed in SW620, and spectral karyotyping identified a complex translocation involving chromosome 19 [53]. There are few reports in the literature investigating associations between loss of this chromosome region and colorectal cancer behavior. In view of the findings presented here, further examination of these two chromosomal regions for loss of heterozygosity or epigenetic silencing in colorectal cancer is warranted.

The bioinformatics approach described here should be of general use for identifying expression within chromosomal domains in both normal and cancerous tissues. It is not surprising that gene expression within chromosomal domains may frequently be disrupted in tumors, given the high frequency with which chromosomes are translocated. This approach is therefore complementary to the commonly used microarray approach which only analyzes gene function. By analyzing both the transcription levels of individual genes and the coordinated expression from within chromosomal domains, a more comprehensive explanation for differential gene expression may be obtained.

Acknowledgments: We thank Junjie Xu for dedicated and excellent technical assistance. This work was supported by the Cancer Society of New Zealand, the Health Research Council, and the Lottery Grant Board.

References and Notes

1. K. W. Kinzler and B. Vogelstein, Lessons from hereditary colorectal cancer, *Cell* 159, 87 **(1996)**.
2. R. Cotran, B. Kumar, and S. Robbins, *Robbins Pathologic Basis of Disease*, Saunders, Philadelphia **(1994)**, 5th Ed., p. 1400.
3. P. Pisani, D. Parkin, F. Bray, and J. Ferlay, Estimates of the worldwide mortality from 25 cancers in 1990, *Int. J. Cancer* 83, 18 **(1999)**.
4. A. B. M. Wilmink, Overview of the epidemiology of colorectal cancer, *Dis. Colon Rectum* 40, 483 **(1997)**.
5. A. Leibovitz, J. C. Stinson, W. B. McCombs, C. E. McCoy, K. C. Mazur, and N. D. Mabry, Classification of human colorectal adenocarcinoma cell lines, *Cancer Res*. 36, 4562 **(1976)**.
6. P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush, A concise guide to cDNA microarray analysis, *Biotechniques* 29, 548 **(2000)**.
7. V. G. Tusher and R. G. C. Tibshirani, Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. U.S.A.* 98, 5116 **(2001)**.
8. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, Systematic determination of genetic network architecture, *Nat. Genet*. 22, 281 **(1999)**.
9. R. Hewitt, K. Brown, and M. Corcoran, Increased expression of tissue inhibitor of metalloproteinases type 1 (TIMP-1) in a more tumorigenic colon cancer cell line, *J. Pathol*. 192, 445 **(2000)**.
10. A. Parle-McDermott, P. McWilliam, O. Tighe, D. Dunican, and D. T. Croke, Serial analysis of gene expression identifies putative metastasis-associated transcripts in colon tumor cell lines, *Br. J. Cancer* 83, 725 **(2000)**.
11. L. H. Butterfield, A. Merino, S. H. Golub, and H. Shau, From cytoprotection to tumor suppression: the multifactorial role of peroxiredoxins, *Antioxid. Redox. Signal.* 1, 385 **(1999)**.
12. S. H. Park, Y. M. Chung, Y. S. Lee, H. J. Kim, J. S. Kim, H. Z. Chae, and Y. D. Yoo, Antisense of human peroxiredoxin II enhances radiation-induced cell death, *Clin. Cancer Res.* 6, 4915 **(2000)**.
13. J. A. Efstathiou, M. Noda, A. Rowan, C. Dixon, R. Chinery, A. Jawhari, T. Hattori, N. A. Wright, W. F. Bodmer, and M. Pignatelli, Intestinal trefoil factor controls the expression of the adenomatous polyposis coli-catenin and the E-cadherin-catenin complexes in human colon carcinoma cells, *Proc. Natl. Acad. Sci. U.S.A.* 95, 3122 **(1998)**.
14. D. R. Taupin, K. Kinoshita, and D. K. Podolsky, Intestinal trefoil factor confers colonic epithelial resistance to apoptosis, *Proc. Natl. Acad. Sci. U.S.A.* 97, 799 **(2000)**.
15. J. Piedra, D. Martinez, J. Castano, S. Miravet, M. Dunach, and A. G. de Herreros, Regulation of beta-catenin structure and activity by tyrosine phosphorylation, *J. Biol. Chem.* 276, 20436 **(2001)**.
16. U. Marklund, N. Larsson, H. M. Gradin, G. Brattsand, and M. Gullberg, Oncoprotein 18 is a phosphorylation-responsive regulator of microtubule dynamics, *EMBO J.* 15, 5290 **(1996)**.
17. G. Brattsand, Correlation of oncoprotein 18/stathmin expression in human breast cancer with established prognostic factors, *Br. J. Cancer* 83, 311 **(2000)**.
18. See http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceSearch.
19. J. H. Kim, M. J. Kang, C. U. Park, H. J. Kwak, Y. Hwang, and G. Y. Koh, Amplified CDK2 and cdc2 activities in primary colorectal carcinoma, *Cancer* 85, 546 **(1999)**.
20. S. Umemura, K. Komaki, S. Noguchi, E. Shiba, M. Toi, I. I. Kimijima, H. Itoh, and Y. R. Osamura, Prognostic factors for node-negative breast cancers: results of a study program by the Japanese breast cancer society, *Breast Cancer* 5, 243 **(1998)**.
21. T. Ohta, M. Fukuda, K. Arima, H. Kawamoto, R. Hashizume, T. Arimura, and S. Yamaguchi, Analysis of Cdc2 and Cyclin D1 expression in breast cancer by immunoblotting, *Breast Cancer* 4, 17 **(1997)**.
22. R. Bravo, Synthesis of the nuclear protein cyclin (PCNA) and its relationship with DNA replication, *Exp. Cell Res.* 163, 287 **(1986)**.
23. Y. Sumiyoshi, Y. Yamashita, T. Maekawa, N. Sakai, T. Shirakusa, and M. Kikuchi, Expression of CD44, vascular endothelial growth factor, and proliferating cell nuclear antigen in severe venous invasional colorectal cancer and its relationship to liver metastasis, *Surg. Today* 30, 323 **(2000)**.
24. G. Liao, Y. Zhang, M. Shen, H. Jiang, and Z. Yan, Significance of proliferating cell nuclear antigen expression in liver metastasis of colorectal cancer, *Dis. Colon Rectum* 40, 1489 **(1997)**.
25. K. Pogue-Geile, J. R. Geiser, M. Shu, C. Miller, I. G. Wool, A. I. Meisler, and J. M. Pipas, Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein, *Mol. Cell Biol.* 11, 3842 **(1991)**.
26. M. G. Denis, C. Chadeneau, M. T. Lecabellec, B. LeMoullac, B. LeMevel, K. Meflah, and P. Lustenberger, Over-expression of the S13 ribosomal protein in actively growing cells, *Int. J. Cancer* 55, 275 **(1993)**.
27. M. Aragona, S. Panetta, A. M. Silipigni, D. L. Romeo, G. Pastura, M. Mesiti, S. Cascinu, and F. La Torre, Nerve growth factor receptor immunoreactivity in breast cancer patients, *Cancer Invest.* 19, 692 **(2001)**.
28. Y. Sakamoto, Y. Kitajima, G. Edakuni, T. Hamamoto, and K. Miyazaki, Combined evaluation of NGF and p75NGFR expression is a biomarker for predicting prognosis in human invasive ductal breast carcinoma, *Oncol. Rep.* 8, 973 **(2001)**.
29. S. Krygier and D. Djakiew, Molecular characterization of the loss of p75(NTR) expression in human prostate tumor cells, *Mol. Carcinog.* 31, 46 **(2001)**.
30. M. D. Benson, S. Eyanson, and N. S. Fineberg, Serum amyloid A in carcinoma of the lung, *Cancer* 57, 1783 **(1986)**.

31. C. Rosenthal and L. Sullivan, Serum amyloid A to monitor cancer dissemination, *Ann. Intern. Med.* 91, 383 **(1979)**.

32. P. S. Weinstein, M. Skinner, J. D. Sipe, J. J. Lokich, N. Zamcheck, and A. S. Cohen, Acute-phase proteins or tumor markers: the role of SAA, SAP, CRP and CEA as indicators of metastasis in a broad spectrum of neoplastic diseases, *Scand. J. Immunol.* 19, 193 **(1984)**.

33. D. K. Trask, V. Band, D. A. Zajchowski, P. Yaswen, T. Suh, and R. Sager, Keratins as markers that distinguish normal and tumor-derived mammary epithelial cells, *Proc. Natl. Acad. Sci. U.S.A.* 87, 2319 **(1990)**.

34. P. Maddox, P. Sasieni, A. Szarewski, M. Anderson, and A. Hanby, Differential expression of keratins 10, 17, and 19 in normal cervical epithelium, cervical intraepithelial neoplasia, and cervical carcinoma, *J. Clin. Pathol.* 52, 41 **(1999)**.

35. M. J. Malecha and M. Miettinen, Expression of keratin 13 in human epithelial neoplasms, *Virchows Arch. A Pathol. Anat. Histopathol.* 418, 249 **(1991)**.

36. L. Nan, H. Kawamata, X. Tan, S. Kameyama, and R. Oyasu, Differential expression of keratin 5 gene in non-tumorigenic and tumorigenic rat bladder cell lines, *Cancer. Lett.* 75, 87 **(1993)**.

37. L. Mishra, B. Bass, B. S. Ooi, A. Sidawy, and L. Korman, Role of insulin-like growth factor-I (IGF-I) receptor, IGF-I, and IGF binding protein-2 in human colorectal cancers, *Growth Horm. IGF Res.* 8, 473 **(1998)**.

38. A. Hoflich, H. Lahm, W. Blum, H. Kolb, and E. Wolf, Insulin-like growth factor-binding protein-2 inhibits proliferation of human embryonic kidney fibroblasts and of IGF-responsive colon carcinoma cell lines, *FEBS Lett.* 434, 329 **(1998)**.

39. N. P. Michell, S. Dent, M. J. Langman, and M. C. Eggo, Insulin-like growth factor binding proteins as mediators of IGF-I effects on colon cancer cell proliferation, *Growth Factors* 14, 269 **(1997)**.

40. K. Swisshelm, K. Ryan, K. Tsuchiya, and R. Sager, Enhanced expression of an insulin growth factor-like binding protein (mac25) in senescent human mammary epithelial cells and induced expression with retinoic acid, *Proc. Natl. Acad. Sci. U.S.A.* 92, 4472 **(1995)**.

41. V. Hwa, C. Tomasini-Sprenger, A. L. Bermejo, R. G. Rosenfeld, and S. R. Plymate, Characterization of insulin-like growth factor-binding protein-related protein-1 in prostate cells, *J. Clin. Endocrinol. Metab.* 83, 4355 **(1998)**.

42. R. J. Cousins and L. Lanningham-Foster, Regulation of cysteine-rich intestinal protein, a zinc finger protein, by mediators of the immune response, *J. Infect. Dis.* 182, S81 (Suppl. 1) **(2000)**.

43. H. Hashida, A. Takabayashi, M. Adachi, T. Imai, K. Kondo, N. Kohno, Y. Yamaoka, and M. Miyake, The novel monoclonal antibody MH8-4 inhibiting cell motility recognizes integrin alpha 3: inverse of its expression withmetastases in colon cancer, *Int. J. Oncol.* 18, 89 **(2001)**.

44. C. Festuccia, G. L. Gravina, A. Angelucci, D. Millimaggi, and M. Bologna, Culture conditions modulate cell phenotype and cause selection of subpopulations in PC3 prostate cancer cell line, *Anticancer Res.* 20, 4367 **(2000)**.

45. N. Pouliot, E. C. Nice, and A. W. Burgess, Laminin-10 mediates basal and EGF-stimulated motility of human colon carcinoma cells via alpha(3)beta(1) and alpha(6)beta(4) integrins, *Exp. Cell Res.* 266, 1 **(2001)**.

46. M. Shang, N. Koshikawa, S. Schenk, and V. Quaranta, The LG3 module of laminin-5 harbors a binding site for integrin alpha3beta1 that promotes cell adhesion, spreading, and migration, *J. Biol. Chem.* 276, 33045 **(2001)**.

47. C. Purdie, J. Piris, C. Bird, and A. Wyllie, 17q allele loss is associated with lymph node metastasis in locally aggressive human colorectal cancer, *J. Pathol*. 175, 297 **(1995)**.

48. C. R. Berney, R. J. Fisher, J. Yang, P. J. Russell, and P. J. Crowe, Genomic alterations (LOH, MI) on chromosome 17q21-23 and prognosis of sporadic colorectal cancer, *Int. J. Cancer.* 89, 1 **(2000)**.

49. E. Campo, R. Miquel, P. Jares, F. Bosch, M. Juan, A. Leone, J. Vives, A. Cardesa, and J. Yague, Prognostic significance of the loss of heterozygosity of Nm23-H1 and p53 genes in human colorectal carcinomas, *Cancer* 73 **(1994)**.

50. B. Iacopetta, S. DiGrandi, B. Dix, C. Haig, R. Soong, and A. House, Loss of heterozygosity of tumor suppressor gene loci in human colorectal carcinoma, *Eur. J. Cancer* 30A, 664 **(1994)**.

51. R. Lamb, J. Going, I. Pickford, and G. Birnie, Allelic imbalance at NME1 in microdissected primary and metastatic human colorectal carcinomas is frequent but not associated with metastasis to lymph nodes or liver, *Cancer Res.* 15, 916 **(1996)**.

52. U. Patel, H. Chen, and S. Banerjee, Dinucleotide repeat polymorphisms at nine loci in sporadic colorectal cancer, *Cell Mol. Biol. Res.* 40, 683 **(1994)**.

53. R. Melcher, C. Steinlein, W. Feichtinger, C. Muller, T. Menzel, H. Luhrs, and W. M. S. Scheppach, Spectral karyotyping of the human colon cancer cell lines SW480 and SW620, *Cytogenet. Cell Genet.* 88, 145 **(2000)**.