Folkesundhed

Screeningsprogrammer for sygdomme baseret på genetiske markører med brug af AI.

Ideen til dette undervisningsforløb stammer fra den vedhæftede artikel om screening for tarmkræft ved brug af fem genetiske markører, som samlet set har vist sig at være tilstrækkeligt forskellige fra syge til raske personer til at kunne danne baggrund for et screeningsprogram.

Artiklen er meget teknisk, men den præsenterer også nogle relevante overvejelser omkring screeningsprogrammer generelt. De behøver ikke være ufejlbarlige på individniveau for at forbedre folkesundheden, men skal fange mange syge tidligt i forløbet, så behandling er mulig. De skal også være enkle og helst ikke invasive, så mange vælger screeningen til.

Vi har ikke tid og ressourcer til at etablere et datagrundløb fra en stor klinisk undersøgelse af mange mennesker, så vi vil kunstigt og med brug af matematik generere et tilsvarende datamateriale, som vi kan bruge som eksempel. Vi vil i den forbindelse bruge de indbyrdes forhold mellem syge og raske for de fem udvalgte genetiske markører fra artiklens tabel 2. Det giver forhåbentlig et realistisk bud på, hvor store forskelle mellem raske og syge, man kan forvente at finde ved genundersøgelser i forbindelse med andre alvorlige sygdomme.

Fra artiklens har vi følgende forhold syge/raske for de fem udvalgte genetiske markører:
Gen1: 0,43
Gen2: 0,42
Gen3: 1,34
Gen4: 1,29
Gen5: 0,42

Bemærk, at nogle i gennemsnit er højere for de syge og nogle er lavere (fra 42% til 134%).

Vi vil generere talværdier for genmarkørerne ved at sample en normalfordeling, hvor vi har valgt en middelværdi som den gennemsnitlige værdi for den pågældende genmarkør og en spredning som et mål for den biologiske variation.

Aktivitet 1 (Afvikles af læreren)

Repeter lidt deskriptiv statistik fra ugrupperede observationer til grupperede observationer og histogrammer. Brug elevernes højder som eksempel.

Bemærk, at histogrammer ofte har en klokkeform, og inddrag flere eksempler blandt andet med afsæt i biologisk variation.

Præsenter tæthedsfunktioner fra normalfordelingen som en model for klokkeformen, som har vist sig at passe godt i mange anvendelser. Sandsynligheder er arealer under grafen.

Vis hvordan man kan sample data med normalfordelingen. Brug f.eks. højder af værnepligtige $N(\mu, \sigma)$ med middelværdi $\mu = 180{,}7\text{cm}$ og spredning $\sigma = 6{,}8\text{cm}$.

For raske har vi valgt at bruge middelværdien $\mu = 10$ og spredningen $\sigma = 5$ altså normalfordelingen $N(10,5)$.

For syge ændrer vi middelværdien afhængig af det angivne forhold for den pågældende genmarkør, men beholder en spredning på 5.
For gen1 bruges $N(4.3,5)$.
For gen2 bruges $N(4.2,5)$.
For gen3 bruges $N(13.4,5)$.
For gen4 bruges $N(12.9,5)$.
For gen5 bruges $N(4.2,5)$.

Vi kan nu generere vores kunstige kliniske data ved at sample disse normalfordelinger.

Det kan gøres på følgende måde i Maple:

Først definerer vi de nødvendige normalfordelinger:

```
with(Statistics):
X := RandomVariable(Normal(10, 5)):
Y1 := RandomVariable(Normal(4.3, 5)):
Y2 := RandomVariable(Normal(4.2, 5)):
Y3 := RandomVariable(Normal(13.4, 5)):
Y4 := RandomVariable(Normal(12.9, 5)):
Y5 := RandomVariable(Normal(4.2, 5)):
```

Vi genererer nu de kliniske data for 10 raske og 10 syge personer. Under preferencer og precision har jeg sat outputtet i Maple til to decimaler. Tallene i hver søjle svarer til talværdierne for gen1 til gen5 for en person:

$$Raske := \begin{bmatrix} Sample(X, 10) \\ Sample(X, 10) \\ Sample(X, 10) \\ Sample(X, 10) \\ Sample(X, 10) \end{bmatrix}:$$

$$Syge := \begin{bmatrix} Sample(Y1, 10) \\ Sample(Y2, 10) \\ Sample(Y3, 10) \\ Sample(Y4, 10) \\ Sample(Y5, 10) \end{bmatrix}:$$

Resultaterne bliver naturligvis forskellige hver gang man sampler normalfordelingerne. Herunder er et eksempel:

$$Raske = \begin{bmatrix} [\ 10.97\ 18.33\ 15.68\ 14.37\ 4.02\ 7.60\ 14.90\ 13.67\ 7.19\ 11.88\ ] \\ [\ 9.64\ 10.30\ 6.94\ 7.63\ 9.45\ 10.17\ 5.61\ -0.63\ 11.75\ 14.24\ ] \\ [\ 13.39\ 13.56\ 2.64\ 13.64\ 12.23\ 7.37\ 15.40\ 12.44\ 3.73\ 10.22\ ] \\ [\ 4.41\ 4.33\ 11.28\ 8.88\ 14.25\ 15.43\ 12.25\ 3.11\ 0.29\ 10.43\ ] \\ [\ 16.89\ 4.39\ 18.06\ 8.99\ 2.24\ 8.74\ 5.72\ 14.47\ 4.63\ 9.27\ ] \end{bmatrix}$$

$$Syge = \begin{bmatrix} [\ 11.57\ 2.32\ 8.60\ 9.85\ 3.46\ 4.93\ -2.64\ 11.82\ 13.40\ 3.88\ ] \\ [\ 3.61\ -0.65\ 0.35\ 9.17\ 3.75\ 0.64\ 1.33\ 4.75\ 5.45\ 3.58\ ] \\ [\ 10.31\ 16.61\ 19.46\ 10.06\ 14.03\ 18.19\ 10.16\ 6.90\ 13.90\ 7.82\ ] \\ [\ 18.10\ 11.76\ 12.35\ 15.12\ 6.61\ 9.81\ 21.48\ 21.03\ 21.36\ 15.37\ ] \\ [\ 2.11\ 6.44\ 2.75\ 6.25\ 4.88\ 1.38\ 9.44\ 0.62\ 1.98\ -0.15\ ] \end{bmatrix}$$

Vi har nu de kliniske data fra vores 20 fiktive personer. F.eks. kan vi aflæse, at for den syge person nummer tre er tallene for genmarkørerne:

$$\begin{bmatrix} gen1 \\ gen2 \\ gen3 \\ gen4 \\ gen5 \end{bmatrix} = \begin{bmatrix} 8.60 \\ 0.35 \\ 19.46 \\ 12.35 \\ 2.75 \end{bmatrix}$$

Aktivitet 2 (Eleverne genererer fiktive kliniske data - skal bruges senere)

Generer kunstige kliniske data for raske og for syge ved at sample normalfordelingerne ovenfor.

Prøv også generere data, hvor spredningen er sat ned fra 5 til 2. Det bør give en større forskel på de kliniske data for syge og for raske.

Prøv tilsvarende at generere data, hvor spredningen er sat op fra 5 til 10. Det bør give en mindre forskel på de kliniske data for syge og for raske.

Bemærk, at vi har valgt at bruge normalfordelingerne $N(10,5)$ som udgangspunkt. Det er blot for at illustrere metoden. Det har altså ikke afsæt i kendte biologiske parametre. At sætte middelværdien til 10 svarer blot til at vælge en enhed, hvor talværdien bliver 10, så det er ret uproblematisk. At sætte spredningen til 5 kan være mere problematisk, da det svarer til at angive information om den biologiske variation for den pågældende genmarkør.

I en virkelig anvendelse med brug af rigtige kliniske data for raske og for syge ville man sikkert se, at normalfordelinger er en god model for de kliniske data, men at middelværdi og spredning er forskellig for hver af de fem genmarkører.

Aktivitet 3 (Afvikles af læreren)

Giv en kort introduktion til træning af neurale net. Gerne med brug af aalborg-intelligence.ai

Vi vil nu illustrere, hvordan de kliniske data kan bruges til at træne et neuralt net, som efterfølgende kan bruges i vores screeningsprogram for den pågældende sygdom. Vi lader de første syv raske og syv syge personer udgøre vores træningsdata.

Først sørger vi for, at Maple kan vise større matricer på skærmen:

```
with(Gym):
visMatrix(25):
```

Vi indlæser også Maples DeepLearning-pakke:

```
with(DeepLearning):
```

Herefter indtaster vi vores træningsdata idet vi tildeler de raske resultatet 1 og de syge resultatet 0:

```
Træningsdata := DataFrame(⟨10.97, 9.64, 13.39, 4.41, 16.89, 1; 18.33, 10.30, 13.56, 4.33, 4.39, 1; 15.68, 6.94,
    2.64, 11.28, 18.06, 1; 14.37, 7.63, 13.64, 8.88, 8.99, 1; 4.02, 9.45, 12.23, 14.25, 2.24, 1; 7.6, 10.17, 7.37,
    15.43, 8.74, 1; 14.9, 5.61, 15.4, 12.25, 5.72, 1; 11.57, 3.61, 10.31, 18.1, 2.11, 0; 2.32, −0.65, 16.61, 11.76,
    6.44, 0; 8.6, 0.35, 19.46, 12.35, 2.75, 0; 9.85, 9.17, 10.06, 15.12, 6.25, 0; 3.46, 3.75, 14.03, 6.61, 4.88, 0; 4.93,
    0.64, 18.19, 9.81, 1.38, 0; −2.64, 1.33, 10.16, 21.48, 9.44, 0⟩, 'rows'= ['person1','person2','person3','person4',
    'person5','person6','person7','person8','person9','person10','person11','person12','person13','person14'],
    'columns'= ['gen1','gen2','gen3','gen4','gen5','resultat']);
```

Det giver følgende tabel

Træningsdata :=

	gen1	gen2	gen3	gen4	gen5	resultat
person1	10.97	9.64	13.39	4.41	16.89	1
person2	18.33	10.30	13.56	4.33	4.39	1
person3	15.68	6.94	2.64	11.28	18.06	1
person4	14.37	7.63	13.64	8.88	8.99	1
person5	4.02	9.45	12.23	14.25	2.24	1
person6	7.6	10.17	7.37	15.43	8.74	1
person7	14.9	5.61	15.4	12.25	5.72	1
person8	11.57	3.61	10.31	18.1	2.11	0
person9	2.32	−0.65	16.61	11.76	6.44	0
person10	8.6	0.35	19.46	12.35	2.75	0
person11	9.85	9.17	10.06	15.12	6.25	0
person12	3.46	3.75	14.03	6.61	4.88	0
person13	4.93	0.64	18.19	9.81	1.38	0
person14	−2.64	1.33	10.16	21.48	9.44	0

Vi kan nu definere og træne et neuralt net med vores kliniske data:

```
model := Sequential([DenseLayer(12, activation = "relu"), DenseLayer(8, activation = "relu"),
    DenseLayer(1, activation = "sigmoid")]):
model:-Compile(loss = "binary_crossentropy", optimizer = "adam", metrics = ["accuracy"]):
model:-Fit(Træningsdata[.., 1..5], Træningsdata[resultat], epochs = 500, batchsize = 14):
```

Vi har nu et neuralt net, som er trænet og klart til brug.

Før vi kan bruge det i et screeningsprogram, skal vi bruge de resterende seks personer fra vores fiktive kliniske data til at teste, om det virker efter hensigten. Vi indtaster:

```
Testdata := DataFrame(⟨13.67, −0.63, 12.44, 3.11, 14.47, 1; 7.19, 11.75, 3.73, 0.29, 4.63, 1; 11.88,
    14.24, 10.22, 10.43, 9.27, 1; 11.82, 4.75, 6.9, 21.03, 0.62, 0; 13.4, 5.45, 13.9, 21.36, 1.98, 0; 3.88,
    3.58, 7.82, 15.37, −0.15, 0⟩, 'rows'= ['personA','personB','personC','personD','personE','personF'],
    'columns'= ['gen1','gen2','gen3','gen4','gen5','resultat']);
```

og det giver

$$Testdata := \begin{bmatrix} & gen1 & gen2 & gen3 & gen4 & gen5 & resultat \\ personA & 13.67 & -0.63 & 12.44 & 3.11 & 14.47 & 1 \\ personB & 7.19 & 11.75 & 3.73 & 0.29 & 4.63 & 1 \\ personC & 11.88 & 14.24 & 10.22 & 10.43 & 9.27 & 1 \\ personD & 11.82 & 4.75 & 6.9 & 21.03 & 0.62 & 0 \\ personE & 13.4 & 5.45 & 13.9 & 21.36 & 1.98 & 0 \\ personF & 3.88 & 3.58 & 7.82 & 15.37 & -0.15 & 0 \end{bmatrix}$$

Vi tester så det neurale net:

```
model:-Evaluate(Testdata[ .., 1 ..5], Testdata[resultat]) :
Testdata[1 ..6, resultat], Modelresultat
```

$$\begin{bmatrix} personA & 1 \\ personB & 1 \\ personC & 1 \\ personD & 0 \\ personE & 0 \\ personF & 0 \end{bmatrix}, \begin{bmatrix} 0.999999046325684 \\ 0.999357640743256 \\ 0.909794747829437 \\ 0.00608994672074914 \\ 0.000143546887557022 \\ 0.844151437282562 \end{bmatrix}$$

Vi ser, at modellen stort set ser ud til at virke, men dog ikke fanger, at personF er syg. PersonD og personE ville imidlertid være blevet opdaget ved en screening.

Kvaliteten af en AI anvendelse afhænger voldsomt af kvaliteten og mængden af træningsdata.

Aktivitet 4 (Eleverne bruger egne fiktive kliniske data til træning af et neuralt net og efterfølgende test af at screeningen med AI virker)

Brug nogle af jeres egne fiktive kliniske data fra aktivitet 2 som træningsdata og andre af jeres kliniske data som testdata.

Definer og træn et neuralt net med jeres træningsdata.

Afprøv derefter med jeres testdata om screeningen for sygdommen med AI virker.

Novel Blood-Based, Five-Gene Biomarker Set for the Detection of Colorectal Cancer

Mark Han,[1] Choong Tsek Liew,[2] Hong Wei Zhang,[1] Samuel Chao,[1] Run Zheng,[1] Kok Thye Yip,[3] Zhen-Ya Song,[4] Hiu Ming Li,[2] Xiao Ping Geng,[6] Li Xin Zhu,[6] Jian-Jiang Lin,[5] K. Wayne Marshall,[1] and Choong Chin Liew[1,7]

Abstract

Purpose: We applied a unique method to identify genes expressed in whole blood that can serve as biomarkers to detect colorectal cancer (CRC).

Experimental Design: Total RNA was isolated from 211 blood samples (110 non-CRC, 101 CRC). Microarray and quantitative real-time PCR were used for biomarker screening and validation, respectively.

Results: From a set of 31 RNA samples (16 CRC, 15 controls), we selected 37 genes from analyzed microarray data that differed significantly between CRC samples and controls ($P < 0.05$). We tested these genes with a second set of 115 samples (58 CRC, 57 controls) using quantitative real-time PCR, validating 17 genes as differentially expressed. Five of these genes were selected for logistic regression analysis, of which two were the most up-regulated (CDA and MGC20553) and three were the most down-regulated (BANK1, BCNP1, and MS4A1) in CRC patients. Logit (P) of the five-gene panel had an area under the curve of 0.88 (95% confidence interval, 0.81-0.94). At a cutoff of logit (P) $>+0.5$ as disease (high risk), <-0.5 as control (low risk), and in between as an intermediate zone, the five-gene biomarker combination yielded a sensitivity of 94% (47 of 50) and a specificity of 77% (33 of 43). The intermediate zone contained 22 samples. We validated the predictive power of these five genes with a novel third set of 92 samples, correctly identifying 88% (30 of 34) of CRC samples and 64% (27 of 42) of non-CRC samples. The intermediate zone contained 16 samples.

Conclusion: Our results indicate that the five-gene biomarker panel can be used as a novel blood-based test for CRC.

The American Cancer Society estimates that in 2007 colorectal cancer (CRC) will be the third leading cause of cancer deaths in men (26,000) and women (26,180). Although current screening tests are effective in early CRC diagnosis, only 39% of cases are diagnosed at a localized stage, as compared with 61% for breast cancer and 74% for bladder cancer (1).

Conventional CRC screening tests include fecal occult blood testing, flexible sigmoidoscopy, double-contrast barium enema X-ray, and colonoscopy (2). Each test has advantages and limitations (3–5). Fecal occult blood testing is noninvasive and relatively inexpensive and is recommended by the American Cancer Society for annual screening (5). Properly done, screening fecal occult blood testing allows CRC to be detected before the onset of symptoms in some patients, thereby reducing mortality by 15% to 33% (6–8). Recent data suggest that current community fecal occult blood screening does not reduce CRC mortality to the extent predicted by randomized, controlled trials (9). A contributing cause seems to be lack of patient compliance; only 20% of U.S. adults, 50 years and older, have annual fecal occult blood testing (2).

Colonoscopy is the "gold standard" for CRC diagnosis and screening and is recommended once every 10 years for average risk persons (2). Colonoscopy is expensive, invasive, frequently not readily available, and occasionally has serious complications. Additionally, many are unwilling to undergo screening colonoscopy: only 29% of average risk persons reported having a colonoscopy in the last decade (10). Importantly, <10% of average risk persons for CRC have advanced adenomas or CRC. Thus, colonoscopy is arguably "overkill" as a primary screening tool. Fecal DNA testing (11) and computed tomography examinations of the colon (virtual colonoscopy; ref. 12) are being investigated; however, their clinical utility is still unclear.

A sensitive, specific, blood-based, noninvasive test for early-stage CRC that does not depend on stool sampling has the

Authors' Affiliations: [1]GeneNews Corporation, Toronto, Canada; [2]Department of Anatomical and Cellular Pathology, Chinese University of Hong Kong, Hong Kong, China; [3]Lam Wah Ee Hospital, Penang, Malaysia; [4]Department of Digestive Medicine, 2nd Affiliated Hospital of Medical College of Zhejiang University; [5]Department of Colorectal Surgery, 1st Affiliated Hospital of Medical College of Zhejiang University, Hangzhou, China; [6]Department of Surgery, 1st Affiliated Hospital, Anhui Medical University, Hefei, China; and [7]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Received 7/20/07; revised 9/11/07; accepted 10/19/07.

Grant support: GeneNews Corporation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Current address for M. Han: Gene Diagnostic Inc., Hangzhou, China.

Requests for reprints: Choong Chin Liew, GeneNews Corporation, 2 East Beaver Creek Road, Building 2, Richmond Hill, Ontario, Canada L4B 2N3. Phone: 617-834-4371; E-mail: cliew@genenews.com.

doi:10.1158/1078-0432.CCR-07-1801

Table 1. Six primer sequences for genes used for quantitative real-time PCR

Symbol	Description	5′ Primer sequences	Position	3′ Primer sequences	Position	Amplicons (bp)
MS4A1	Membrane-spanning 4-domains, subfamily A, member 1	AAAGAACGTGCTCCAGACCC	970	TTCAGTTAGCCCAACCACTTCTTC	1,075	106
CDA	Cytidine deaminase	ATCGCCAGTGACATGCAAGA	376	TACCATCCGGCTTGGTCAGTA	484	109
BCNP1	B-cell novel protein 1	GGCGCGTGCTGAAGAAATTCAA	1,664	TTTGCAGCCTGGCTCGAGTTG	1,782	119
MGC20553	FRMD domain containing 3	AGGRGCACAGAGCCAACATTAC	1,268	AATGGAACACCCTCACCCAGA	1,413	146
BANK1	B-Cell scaffold protein with ankyrin repeats 1	TGCTGAAAGGCATGGTCACAAAG	1,303	GCTGGGTTCTGTGTGGAAGGAATA	1,449	147
ACTB	Actin, β	TCAAGATCATTGCTCCTCCTGAGC	1,064	AAGCATTTGCGGTGGACGAT	1,208	145

potential for greater patient compliance, with associated public health benefits and decreased health care costs.

There is evidence that unique gene expression patterns in peripheral blood reflect static (inherited) and dynamic (environmental) changes that occur within the cells and tissues of the body (Sentinel Principle; ref. 13). Our laboratory has applied the Sentinel Principle across a broad range of diseases, including schizophrenia, cardiovascular disease, osteoarthritis, and bladder cancer (14–17). A preliminary report indicates that a blood biomarker panel was able to detect human CRC (18). Other laboratories have independently shown that blood gene expression can differentiate disease samples from controls (19–22).

In this study, we profiled peripheral blood samples from CRC patients and non-CRC controls. We validated a group of CRC transcript biomarkers using quantitative real-time PCR. We yielded a best logit (*P*) equation from five selected genes to classify CRC from non-CRC and tested the predictive performance of this five-gene biomarker panel against an independent set of test samples.

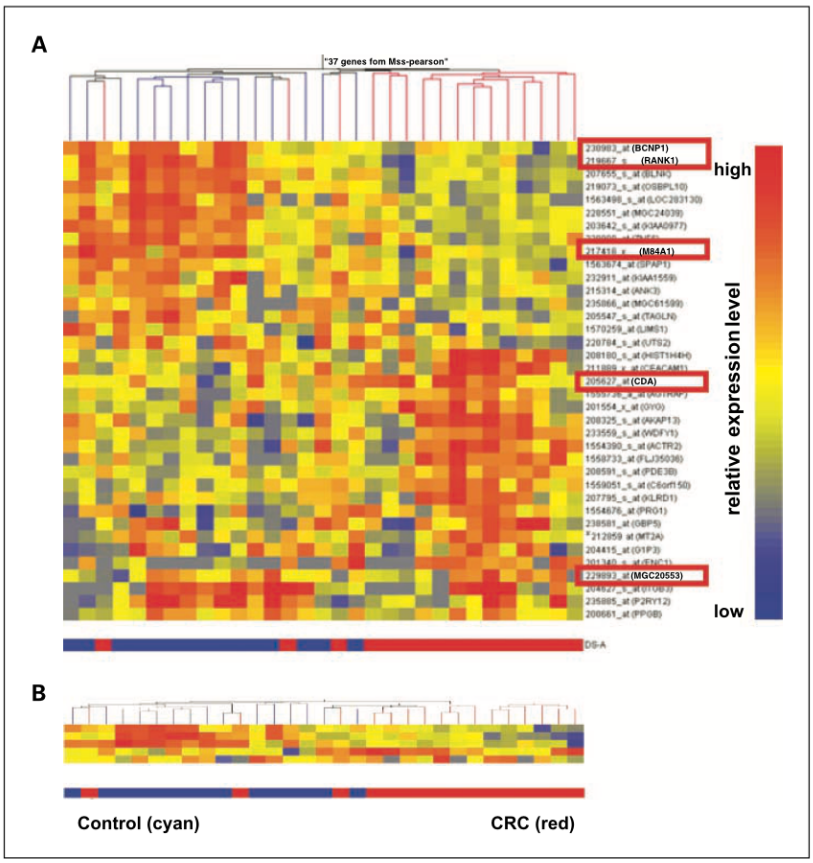

Fig. 1. Hierarchical cluster analysis of identified probe sets of microarray data. Thirty-seven probe sets (*A*) and five probe sets (*B*) highlighted in red box in A were differentially expressed in blood between controls and CRC patients. Cyan, control ($n = 15$); red, CRC ($n = 16$). *212859 at (MT2A) was formerly MT1E.

Materials and Methods

Patient samples. We obtained blood samples from 211 subjects (110 non-CRC controls, 101 CRC patients) recruited from four institutions. CRC blood samples were collected before tumor resection; cancer stage and histology were determined by institutional pathologists. Of the 101 CRC patients, 77 were Dukes stages: A, 8; B, 31; C, 31; and D, 7. The controls comprised individuals from the same institutions, including noncancer patients and healthy volunteers. Informed consent was obtained according to the research protocols approved by the research ethical boards of the institutions involved.

A 31-sample set (15 controls, 16 CRC) was used for gene profiling on Affymetrix U133Plus 2.0 chip (Affymetrix). A second set of 115 samples (data set AL8a1, 88 additional samples plus 27 samples for Affymetrix) composed of 57 controls and 58 CRC samples was a training set for quantitative real-time PCR validation. A third set of independent 92 samples (data set AL8a2, 49 controls and 43 CRC) was a blind test set.

Blood collection and RNA isolation. Samples of peripheral whole blood (10 mL) were collected in EDTA Vacutainer tubes (Becton Dickinson) and stored at 4°C until processing (within 6 h). Plasma was removed after centrifugation and a hypotonic buffer (1.6 mmol/L EDTA, 10 mmol/L $KHCO_3$, 153 mmol/L NH_4Cl, pH 7.4) was added at a 3:1 volume ratio to lyse the RBC. The mixture was centrifuged to yield a pellet containing WBC, and the pellet was resuspended into 1.0 mL of TRIzol reagent (Invitrogen Corp.) and 0.2 mL of chloroform. RNA quality was assessed on an Agilent 2100 Bioanalyzer RNA 6000 Nano Chip. RNA quantity was determined by absorbance at 260 nm in a Beckman-Coulter DU640 Spectrophotometer.

Microarray hybridization. Five micrograms of purified total RNA were labeled and hybridized against Affymetrix U133Plus 2.0 GeneChip oligonucleotide arrays (Affymetrix). Hybridization signals were adjusted in the Affymetrix GCOS software (version 1.1.1) using a scaling factor that adjusted the global trimmed mean signal intensity value to 500 for each array and imported into GeneSpring version 6.2 (Silicon Genetics). Signal intensities were centered to the 50th percentile of each chip, and for each individual probe set, to the median intensity of each specific subset first, to minimize possible technical bias, then to the whole sample set. Only genes identified by the GCOS software as "present" or "marginal" in all samples were analyzed.

Quantitative real-time PCR. Two micrograms of RNA was reversed transcribed into single-stranded cDNA using the High-Capacity cDNA Archive Kit (Applied Biosystems) in 100-μL reaction. For training set samples, 2-ng cDNA was mixed with SYBR Green master mix (SYBR Green PCR Kit, Qiagen) and primers in a 20-μL reaction volume. PCR amplification was done using the DNA Engine Opticon (Bio-Rad; formerly MJ Research). For blind test samples, 5-ng cDNA was mixed with the master mix and primers in 25-μL reaction volume. PCR amplification was done on ABI 7500 Real-time PCR System (Applied Biosystems). Dissociation curves generated at the end of each run were examined to verify specific PCR amplification and absence of primer-dimer formation. Forward and reverse primers for target genes were designed using PrimerQuest[8] (Integrated DNA Technologies). The sequences of the six primer sets are listed (Table 1). Amplification efficiency for each primer pair was determined using a serial dilution of reference cDNA generated from a normal blood RNA pool to ensure that values were within linear range, and amplification efficiency was approximately equal for each gene tested. Amplification specificity was confirmed by agarose gel electrophoresis of the PCR products.

Data analysis. Differentially expressed genes were identified by statistically significant differences ($P < 0.05$) in microarray probe set expression values between the CRC and control groups using the Wilcoxon-Mann-Whitney nonparametric test. Supervised hierarchical cluster analysis was done on each comparison to assess correlations among samples for identified probe sets.

For quantitative real-time PCR results, we used the comparative Ct equation (User Bulletin #2, Applied Biosystems, 2001) to calculate relative fold changes (CRC versus controls). Welch t test was used to evaluate the differences in mRNA levels between controls and CRC patients. Differences were considered significant when $P < 0.05$.

We used logistic regression to analyze the dependence of the binary diagnostic variable Y (control, 0; disease, 1) on the ΔCt values from the training data set. When P (probability of a patient sample) is diagnosed as "diseased," then a function $X = \text{logit}(P)$ can be defined as follows:

$$X = \text{logit}(P) = \ln(P/(1-P)) = b_0 + b_1 \Delta\text{Ct}_1 + b_2 \Delta\text{Ct}_2 + \ldots + b_n \Delta\text{Ct}_n \quad \text{(A)}$$

Maximum-likelihood fitting method was used to obtain the (empirical) coefficients $\{b_i\}$ that define the relationship between X and the experimental measurements $\{\Delta\text{Ct}_i\}$. The $\{b_i\}$ values were obtained using MedCalc software program (MedCalc Software). Receiving operating characteristic (ROC) curve analysis was then used to evaluate the discriminatory power of the combinations (23). Classification power was determined by area under the curve (AUC), sensitivity, and specificity at the defined cutoff.

For cross-validation, data (ΔCt values) were analyzed using the "simple logistic regression" function of WEKA[9] under "Experimenter" mode. We applied 5-fold (where 4/5 of the samples as training set and the remainder 1/5 of the samples as test set) "cross validation" with iteration at a number of 1,000 repetitions. The output file is further analyzed using Excel to calculate the average accuracy for all iterations.

For prediction test, a blind test set was examined against the five genes. The ΔCt values were used to calculate logit function X_i using the coefficients defined from the training set (Eq. A).

Results

Identification of differentially expressed genes using microarrays. To screen differentially expressed genes between CRC and non-CRC samples, we hybridized a 31-sample set using Affymetrix Gene Chips. We selected 37 probe sets not related to age or gender, significantly different in blood gene expression between controls and CRC ($P < 0.05$). The 37 genes were selected by P value, fold change, and gene function and used for hierarchical cluster analysis to identify sample similarities in expression patterns, as previously described (24). Analysis of these differentially expressed genes resulted in a separation of cancer group from controls (Fig. 1A).

Validation of CRC biomarkers using quantitative real-time PCR. We tested the 37 selected genes with the training set of 115 samples using SYBR Green quantitative real-time PCR. Seventeen genes were statistically significantly different ($P < 0.05$) in the two groups (Table 2). Fourteen of these genes were significantly underexpressed and three were overexpressed in CRC patients as compared with controls. The two most significantly up-regulated genes and the three most down-regulated genes are highlighted in Table 2. Cytidine deaminase (*CDA*) and FERM domain containing 3 (*MGC20553*) were expressed 1.34-fold ($P = 0.0001$) and 1.29-fold ($P = 0.0232$) higher in CRC patients (red; Table 2); and B-cell novel protein 1 (*BCNP1*), B-cell scaffold protein with ankyrin repeats 1 (*BANK1*), and membrane-spanning 4 domains, subfamily A, member 1 (*MS4A1*) were expressed 0.42-fold ($P < 0.0001$),

[8] http://biotools.idtdna.com/primerquest

[9] http://www.cs.waikato.ac.nz/ml/weka/

Table 2. Fold change of quantitative real-time PCR result and AUC of the five genes used for logistic regression

Gene	P	Fold (cancer/control)	AUC (95% CI)
BANK1	0.0000	0.43	0.86 ± 0.04 (0.78-0.92)
BCNP1	0.0000	0.42	0.84 ± 0.04 (0.76-0.90)
CDA	0.0001	1.34	0.70 ± 0.05 (0.61-0.78)
MGC20553	0.0232	1.29	0.62 ± 0.05 (0.52-0.71)
MS4A	0.0000	0.42	0.82 ± 0.04 (0.74-0.88)
ACTR2	0.0331	0.89	
AKAP13	0.0003	0.80	
BLNK	0.0000	0.51	
C9orf85	0.0000	0.65	
COBLL1	0.0000	0.47	
GBP5	0.0324	1.32	
HIST1H4H	0.0000	0.67	
KIAA1559	0.0000	0.61	
LOC283130	0.0000	0.63	
OSBPL10	0.0000	0.45	
SPAP1	0.0000	0.40	
WDFY1	0.0092	0.85	

NOTE: In boldface are the five genes used for combination analysis including three most significantly down-regulated genes and two most significantly up-regulated genes.
Abbreviations: BANK1, B-cell scaffold protein with ankyrin repeats 1; BCNP1, B-cell novel protein 1; CDA, cytidine deaminase; MGC20553, FERM domain containing 3; MS4A1, membrane-spanning 4 domains, subfamily A, member 1; ACTR2, actin-related protein 2; AKAP13, a kinase (PRKA) anchor protein 13; BLNK, B-cell linker; C9orf85, chromosome 9 open reading frame 85; COBLL1, COBL-like 1; GBP5, guanylate binding protein 5; HIST1H4H, histone cluster 1, H4H; KIAA1559, zinc finger protein 14 homologue (mouse); LOC283130, solute carrier family 25, member 45 (SLC25A45); OSBPL10, oxysterol binding protein-like 10; SPAP1, SH2 domain containing phosphatase anchor protein 1 (predicted); WDFY1, WD repeat and FYVE domain containing 1. 95% CI, 95% confidence interval.

0.43-fold ($P < 0.0001$), and 0.42-fold ($P < 0.0001$) lower in CRC patients (blue; Table 2). Hierarchical cluster analysis using these five genes with the microarray data showed a similar separation of the cancer group from the controls (Fig. 1B) as the separation of 37 genes (Fig. 1A).

Discriminatory power of combined biomarkers. ROC analysis of quantitative real-time PCR data for each of the five genes resulted in AUC values between 0.62 and 0.86 (Table 2). To evaluate the discriminatory power of the biomarker combination, we carried out logistic regression analysis for the ΔCt values of five differentially expressed genes. The best combination equation can be formulated as follows: logit (P) = -5.963 + 1.206 × BANK1 ΔCt + 0.879 × BCNP1 ΔCt - 0.881 × CDA ΔCt - 0.375 × MGC20553 ΔCt - 0.405 × MS4A1 ΔCt. The AUC for the five-gene biomarker panel was 0.88 ± 0.03 (95% confidence interval, 0.81-0.94; $P < 0.001$; Fig. 2). Cross-validation (5-fold) of this training set showed an average accuracy of 79% (SD, 7.5%), only 3 percentage points lower than for the training set itself.

Our experience with the quantitative PCR instrument has shown that there is an average SD of ~0.3 cycle on average for the genes in the study.[10] The value of 0.3 cycle was propagated through the logistic regression equation and gave an estimated "gray zone" extending from logit (P) = -0.5 to +0.5. Samples with logit (P) value above the upper threshold of +0.5 are classified as "high risk for cancer," whereas samples with logit (P) value below the lower threshold of -0.5 are classified as "low risk for cancer." Samples with logit (P) values between -0.5 to +0.5 were classified as "intermediate risk" (Fig. 3).

For the training set (n = 115), the five-gene panel yielded a sensitivity of 94% (47 of 50) and a specificity of 77% (33 of 43) with 19% (22 samples) of the samples classified as "intermediate risk" (gray zone). The positive predictive value is 82% and the negative predictive value is 92% (Table 3).

Predictive performance of five-gene biomarker panel. To test the predictive performance of the five-gene biomarker panel generated from the training set, we quantitated mRNA levels for the same five genes by quantitative real-time PCR using 92 independent RNA samples collected from four different sites. This testing resulted in 88% (30 of 34) correct prediction for CRC patients as "high risk," 64% (27 of 42) correct prediction for non-CRC patients as "low risk," and 16 "intermediate risk" predictions. The corresponding positive predictive value is 67%; negative predictive value is 87% (Table 3).

Discussion

In this study, we identified and validated a number of differentially expressed genes between subjects with and without CRC. We reported that a five-gene combination differentiated the two groups with sensitivity and specificity of 94% and 77%, respectively. This five-gene panel represents a novel biomarker set for CRC detection.

The uniqueness of our approach is our use of combinations of biomarkers assayed in whole blood (13, 17). By combining biomarkers, we can obtain higher levels of discrimination and reproducibility than possible with any single biomarker. The AUC from ROC analysis ranges from 0.62 to 0.86 (SD, 0.04-0.05) for each of the five biomarkers individually; the biomarker combination improves diagnostic capability to an AUC of 0.88 (SD, 0.030).

Once the AUC is established for a biomarker panel, the relative trade-off in performance between sensitivity and specificity is determined by the choice of cutoff point. For example, in our training set, if a single cutoff point is used, we can achieve a sensitivity of 95% with a specificity of 58% at a cutoff threshold of -0.5; a sensitivity of 81% with a specificity of 83% at a cutoff threshold of +0.5; and a sensitivity of 90% with a specificity of 79% at a cutoff threshold of 0.0 (Fig. 3).

In this report, we apply the concept of including a gray or intermediate zone to the interpretation of biomarker set results. Due to the technical limitations of quantitative real-time PCR as well as biological variability, an area of overlap occurs in the distribution between the high-risk and low-risk populations. Segregating this intermediate zone from the high-risk and low-risk zones improves the predictive performance of the test for the samples that fall into the high-risk or low-risk category (~80% in total; Table 3; Fig. 3).

This biomarker combination compares favorably in accuracy with fecal occult blood testing and with the fecal DNA test (11). Stool-based tests have relatively low sensitivity (5-25%) and relatively high specificity (80-95%; refs. 11, 25); our biomarker panel has similar specificity but much higher sensitivity.

[10] Unpublished data.

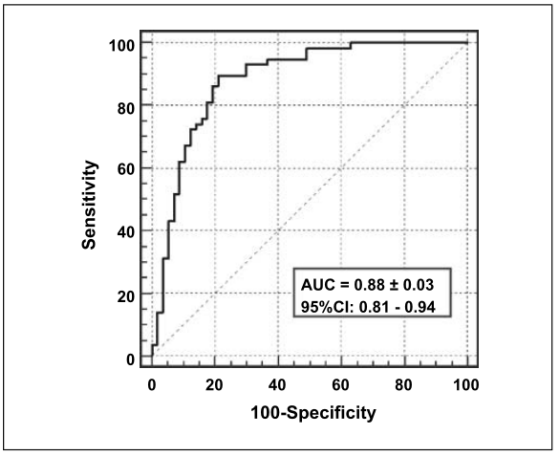

Fig. 2. Discrimination power of five-gene combination. ROC analysis gave an AUC of 0.88. Control, $n = 57$; cancer, $n = 58$.

Table 3. Summary of the five-gene panel performance on training and test sets

CRC	Training set	Test set
True negative	33	27
False positive	10	15
Control (low risk)	43	42
False negative	3	4
True positive	47	30
CRC (high risk)	50	34
Subtotal	93	76
Intermediate risk	22	16
Total	115	92
Sensitivity	94%	88%
Specificity	77%	64%
PPV	82%	67%
NPV	92%	87%

Abbreviations: PPV, positive predictive value; NPV, negative predictive value.

Furthermore, because we use a peripheral blood sample obtained by routine venipuncture, our approach has the advantage over fecal tests of being much more acceptable to patients.

Another advantage of our test is that it has a continuous-valued output: the logit (P). This makes it possible to define a movable threshold, which can be set to achieve a combination of sensitivity and specificity values (from the ROC curve) that best fits the intended use of the test. For example, if the test is intended to be applied in an average-risk population to identify patients who would likely benefit from colonoscopy examination, then the threshold is set for high sensitivity (true positive fraction). By contrast, tests such as fecal occult blood testing, fecal DNA test, and detection of circulating cancer cells in peripheral blood (26) have discrete (yes/no) outputs only.

In this study, we observed increased transcript levels of the gene cytidine deaminase (*CDA*; localized to 1p35-36.2) in blood from CRC patients. CDA is a salvage pathway enzyme that converts cytosine arbinoside (Ara-C) to Ara-U, thereby decreasing the formation of cytosine arabinoside triphosphate (Ara-CTP; 27, 28). Ara-C, a deoxycytidine analogue, is phosphorylated into its active form, Ara-CTP, which competes with dCTP for incorporation into DNA. Incorporated Ara-C blocks DNA synthesis and the cell undergoes programmed cell death. Studies of acute myeloid leukemia in children with and without Down syndrome indicate that elevated CDA transcript levels correlate with poor outcome in Ara-C–based chemotherapy (29, 30). CDA gene expression/activity and outcome of gemcitabine-based treatment also correlate in neuroblastoma cell lines (31) and pancreatic cancer (32). Ara-C and gemcitabine have been used in CRC treatment (33–36), but the correlation between CDA and treatment effectiveness has not been studied. That CDA was overexpressed in the present study suggests potential effects of CDA in Ara-C– or gemcitabine-based CRC treatment and warrants further investigation.

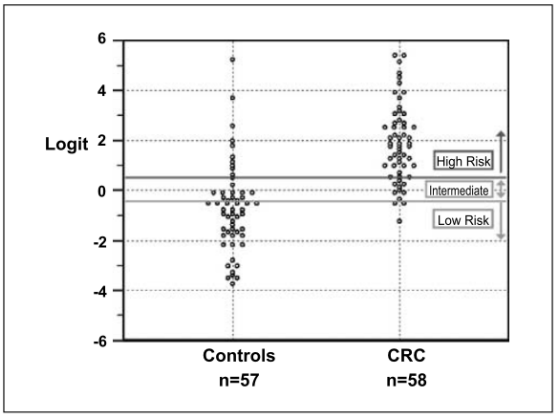

Fig. 3. Sample distribution of the training set. A cutoff >+0.5 is high risk or disease, <-0.5 is low risk or control, and in between is intermediate risk.

MGC20553, also up-regulated in CRC, was initially identified as a novel gene on chromosome 9q22.2-31.1. MGC20553 is a multifunctional protein essential for maintaining erythrocyte shape and membrane mechanical properties (37). The exact function of this gene in blood cells has yet to be determined. MGC20553 was studied in acute myeloid leukemia and no change in its expression was observed (38). Our study showed that MGC20553 helps discriminate between CRC and non-CRC, indicating that MGC20553 is a CRC response gene.

Three genes down-regulated in CRC blood samples, *BCNP1*, *BANK1*, and *MS4A1*, are expressed in B cells. BCNP1 protein, initially identified in chronic lymphocytic leukemia and in B-cell malignancies (39), had three predicted transmembrane domains and no known function. BANK1 is a novel substrate of tyrosine kinases. It is tyrosine phosphorylated on B-cell antigen receptor stimulation, which is mediated predominantly by tyrosine kinase Syk. Overexpression of BANK1 in B cells enhances B-cell antigen receptor–induced calcium mobilization and may be specific to antigen-induced immune responses (40). Gene MS4A1, a member of the membrane-spanning 4A gene family, encodes a B-cell surface molecule that functions

in the differentiation of B-cells into plasma cells (41). Our findings indicate that these three genes might be functionally involved in CRC.

We have identified a peripheral blood biomarker panel able to discriminate CRC from non-CRC samples. This blood-based test will be valuable for screening populations for CRC. The currently recommended fecal occult blood test has relatively high specificity but low sensitivity. One recent study by Imperiale et al. (11) showed a sensitivity of 12.9% and a specificity of 94.4% for fecal occult blood testing in an average-risk screening population. Our blood CRC biomarkers showed a much higher sensitivity than fecal occult blood testing. More importantly, a blood-based test will have much better patient compliance than a fecal-based screening test. The increased rates of compliance expected for a blood test relative to other CRC screening modalities will potentially result in earlier cancer detection, with decreased morbidity and mortality and more effective utilization of health care resources. Further work is ongoing to identify additional specific markers informative for detecting CRC; to refine the algorithm for choosing optimal combinations of markers to be incorporated into a CRC biomarker panel; to examine the CRC biomarker panel using samples from patients with cancers other than CRC; and to examine CRC biomarker panel performance across a larger population.

Acknowledgments

We thank Larry Heisler, Daniel H. Farkas, and Sapna Syngal for their critical comments, and Gerard Beltran, Michael Jones, Su Zhang, and Mary Kazijas for their technical assistance.

References

1. Jemal A, Siegel R, Ward E, et al. Cancer statistics 2007. CA Cancer J Clin 2007;57:43–66.
2. Smith RA, Cokkinides V, Eyre HJ. American Cancer Society guidelines for the early detection of cancer, 2006. CA Cancer J Clin 2006;56:11–25.
3. Muller O. Identification of colon cancer patients by molecular diagnosis. Dig Dis 2003;21:315–9.
4. Ransohoff DF. Colon cancer screening in 2005: status and challenges. Gastroenterology 2005;128: 1685–95.
5. Greenwald B. A comparison of three stool tests for colorectal cancer screening. Medsurg Nurs 2005;14: 292–9.
6. Mandel JS, Church TR, Ederer F, et al. Colorectal cancer mortality: effectiveness of biennial screening for fecal occult blood. J Natl Cancer Inst 1999;91:434–7.
7. Hardcastle JD, Chamberlain JO, Robinson MH, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. Lancet 1996;348: 1472–7.
8. Jorgensen OD, Kronborg O, Fenger C. A randomised study of screening for colorectal cancer using faecal occult blood testing: results after 13 years and seven biennial screening rounds. Gut 2002;50:29–32.
9. Nadel MR, Shapiro JA, Klabunde CN, et al. A national survey of primary care physicians' methods for screening for fecal occult blood. Ann Intern Med 2005;142: 86–94.
10. Wee CC, McCarthy EP, Phillips RS. Factors associated with colon cancer screening: the role of patient factors and physician counseling. Prev Med 2005;41: 23–9.
11. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. N Engl J Med 2004;351:2704–14.
12. Duff SE, Murray D, Rate AJ, et al. Computed tomographic colonography (CTC) performance: one-year clinical follow-up. Clin Radiol 2006;61:932–6.
13. Liew CC, Ma J, Tang HC, et al. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J Lab Clin Med 2006; 147:126–32.
14. Ma J, Liew CC. Gene profiling identifies secreted protein transcripts from peripheral blood cells in coronary artery disease. J Mol Cell Cardiol 2003;35:993–8.
15. Tsuang MT, Nossova N, Yager T, et al. Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: a preliminary report. Am J Med Genet B Neuropsychiatr Genet 2005;133:1–5.
16. Marshall KW, Zhang H, Yager TD, et al. Blood-based biomarkers for detecting mild osteoarthritis in the human knee. Osteoarthritis Cartilage 2005;13:861–71.
17. Osman I, Bajorin DF, Sun TT, et al. Novel blood biomarkers of human urinary bladder cancer. Clin Cancer Res 2006;12:3374–80.
18. Han M, Liew CT, Zhang HW, et al. Novel blood biomarker panel detects human colorectal cancer. Journal of Clinical Oncology 2006 ASCO Annual Meeting Proceedings (Post-Meeting Edition) 2006;24:18S:3611.
19. Whistler T, Unger ER, Nisenbaum R, et al. Integration of gene expression, clinical, and epidemiologic data to characterize chronic fatigue syndrome. J Transl Med 2003;1:10.
20. Bennett L, Palucka AK, Arce E, et al. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med 2003;197:711–23.
21. Tang Y, Gilbert DL, Glauser TA, et al. Blood gene expression profiling of neurologic diseases: a pilot microarray study. Arch Neurol 2005;62:210–5.
22. Deng MC, Eisen HJ, Mehra MR, et al. Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. Am J Transplant 2006;6:150–60.
23. Pampel FC. Logistic regression: a primer. Thousand Oaks (CA): Sage Publications; 2000.
24. Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.
25. Bampton PA, Sandford JJ, Cole SR, et al. Interval faecal occult blood testing in a colonoscopy based screening programme detects additional pathology. Gut 2005;54:803–6.
26. Yeh CS, Wang JY, Wu CH, et al. Molecular detection of circulating cancer cells in the peripheral blood of patients with colorectal cancer by using membrane array with a multiple mRNA marker panel. Int J Oncol 2006;28:411–20.
27. Grant S. Ara-C: cellular and molecular pharmacology. Adv Cancer Res 1998;72:197–233.
28. Galmarini CM, Graham K, Thomas X, et al. Expression of high K_m 5′-nucleotidase in leukemic blasts is an independent prognostic factor in adults with acute myeloid leukemia. Blood 2001;98:1922–6.
29. Taub JW. Relationship of chromosome 21 and acute leukemia in children with Down syndrome. J Pediatr Hematol Oncol 2001;23:175–8.
30. Ge Y, Jensen TL, Stout ML, et al. The role of cytidine deaminase and GATA1 mutations in the increased cytosine arabinoside sensitivity of Down syndrome myeloblasts and leukemia cell lines. Cancer Res 2004;64:728–35.
31. Ogawa M, Hori H, Ohta T, et al. Sensitivity to gemcitabine and its metabolizing enzymes in neuroblastoma. Clin Cancer Res 2005;11:3485–93.
32. Bengala C, Guarneri V, Giovannetti E, et al. Prolonged fixed dose rate infusion of gemcitabine with autologous haemopoietic support in advanced pancreatic adenocarcinoma. Br J Cancer 2005;93: 35–40.
33. Tetef M, Leong L, Ahn C, et al. Cisplatin and infusional cytosine arabinoside for the treatment of colorectal adenocarcinoma: a phase II trial. Cancer Invest 1999;17:114–7.
34. Cha MC, Lin A, Meckling KA. Low dose docosahexaenoic acid protects normal colonic epithelial cells from araC toxicity. BMC Pharmacol 2005;5:7.
35. Saiko P, Horvath Z, Bauer W, et al. *In vitro* and *in vivo* antitumor activity of novel amphiphilic dimers consisting of 5-fluorodeoxyuridine and arabinofuranosylcytosine. Int J Oncol 2004;25:357–64.
36. Hochster HS. The role of pemetrexed in the treatment of gastrointestinal malignancy. Clin Colorectal Cancer 2004;4:190–5.
37. Ni X, Ji C, Cao G, et al. Molecular cloning and characterization of the protein 4.1O gene, a novel member of the protein 4.1 family with focal expression in ovary. J Hum Genet 2003;48:101–6.
38. Sweetser DA, Peniket AJ, Haaland C, et al. Delineation of the minimal commonly deleted segment and identification of candidate tumor-suppressor genes in del(9q) acute myeloid leukemia. Genes Chromosomes Cancer 2005;44:279–91.
39. Boyd RS, Adam PJ, Patel S, et al. Proteomic analysis of the cell-surface membrane in chronic lymphocytic leukemia: identification of two novel proteins, BCNP1 and MIG2B. Leukemia 2003;17:1605–12.
40. Yokoyama K, Su Ih IH, Tezuka T, et al. BANK regulates BCR-induced calcium mobilization by promoting tyrosine phosphorylation of IP(3) receptor. EMBO J 2002;21:83–92.
41. Tedder TF, Disteche CM, Louie E, et al. The gene that encodes the human CD20 (B1) differentiation antigen is located on chromosome 11 near the t(11;14)(q13;q32) translocation site. J Immunol 1989; 142:2555–9.