Cancer Detection and Prevention. 2008; 32(3): 236-50

Bioinformatics approach to mRNA markers discovery for detection of circulating tumor cells in patients with gastrointestinal cancer

Manuel Valladares-Ayerbes, MD[a], Silvia Díaz-Prado, PhD[b], Marga Reboredo, MD[a], Vanessa Medina, Bs, Sc[a], Pilar Iglesias-Díaz, MD[c], Maria J. Lorenzo-Patiño, MD[c], Rosario G. Campelo, MD[a], Mar Haz, Tch[a], Isabel Santamarina, Tch[a], Luis M. Antón-Aparicio, MD, PhD[a]

[a] *Medical Oncology and Institute for Biomedical Research (INIBIC), Juan Canalejo University Hospital, Spain*
[b] *Medicine Department, La Coruña University, Spain*
[c] *Pathology Department, Juan Canalejo University Hospital, Spain*

Abstract
Background: Detection of tumor cells in the blood, or minimal deposits in distant organs as bone marrow, could be important to identify cancer patients at high risk of relapse or disease progression. Quantitative polymerase chain reaction (PCR) amplification of tissue or tumor selective mRNA is the most powerful tool for the detection of this circulating or occult metastatic cells. Our study aims to identify novel gastrointestinal cancer-specific markers for circulating tumor cell detection. *Method*: Phase I preclinical study was performed by means of computational tools for expression analysis. *In silico* data were used to identify and prioritize molecular markers highly expressed in gastrointestinal cancers but absent in hematopoietic-derived libraries. Selected genes were evaluated by means of qRT-PCR in gastrointestinal cancer and hematopoietic cell-lines, normal human bone marrows and bloods, tumor tissue, and blood from cancer patients. *Results*: Novel and known mRNA markers for circulating tumor cell detection in gastrointestinal cancer have been identified. Among all the genes assessed, PKP3, AGR2, S100A16, S100A6, LGALS4, and CLDN3 were selected and assays based on blood qRT-PCR were developed. Reliably qRT-PCR assays for the novel targets plakophilin 3 (PKP3) and anterior gradient-2 (AGR2) to identify blood-borne cells in cancer patients were developed. *Conclusions*: Novel and known gastrointestinal-specific mRNA markers for circulating tumor cells have been identified through *in silico* analysis and validated in clinical material. qRT-PCR assay targeted to PKP3 and AGR2 mRNAs might be helpful to detect circulating tumor cells in patients with gastrointestinal cancer.

Keywords
Neoplasm circulating cells; Metastasis; Biological tumor markers; Reverse transcriptase polymerase chain reaction; Gene expression pattern analysis; Databases nucleic acid; Epithelial cells; Tumor cell lines; Anterior gradient 2 homolog (*Xenopus laevis*) protein; Plakophilin 3

1. Introduction

Cancer of the gastrointestinal (GI) system encompasses a great proportion of tumor-related deaths in the world. Stage at the time of diagnostic and suitability for curative surgery remains the most important prognostic factors for digestive tract tumors. However, distant and loco-regional relapses frequently occur in spite of resection and adjuvant therapy. Colorectal cancer is the second leading cause of cancer deaths in the western countries. Most of these patients are diagnosed with localized disease, but their average 5-year survival is 50–60% in European countries [1]. For patients with other common GI malignancies, the options for curative resections are limited and survival remains poor, especially for esophageal and gastric carcinomas and pancreatic cancer.

Circulating tumor cells and occult metastasis (micrometastasis) are considered early stages in the progression of metastatic cascade in epithelial cancer. Detection of tumor cells in the blood or minimal deposits in distant organs as bone marrow could be important to identify patients at high risk of relapse or disease progression.

PCR amplification of tissue or tumor-selective mRNA is the most powerful tool for detection of this circulating or occult metastatic cells. Cytokeratin 20, carcinoembryonic antigen (CEA), and epidermal growth factor receptor (EGFR) are among the most frequent mRNA markers used in different reverse-transcriptase polymerase-chain reaction (RT-PCR) assays in GI-cancer patients [2] ; [3]. However, cancer genetic heterogeneity, down-regulation of mRNA marker in tumor cells, or low-level transcription of selected target in the hematopoietic compartment could compromise both sensitivity and specificity of molecular methods.

Selection of novel gastrointestinal cancer-specific transcripts and development of multi-marker quantitative RT-PCR assays are clearly outstanding research questions [4] ; [5].

EGFR has been found to be overexpressed in a high proportion of gastrointestinal cancer and confers a poor prognosis [6]. Development of a quantitative assay to detect EGFR-expressing circulating tumor cells could be relevant for prognostic assessment, monitor response, or even thought select target therapy [7].

RT-PCR assays for different glycosyltransferases have been developed for tumor cells detection in blood for breast cancer and melanoma patients [8]. Beta-1,6-*N*-acetylglucosaminyltransferase V (GNT-V, MGAT5; Hs.200/Mm.161; EC 4.1.155) expression in colorectal cancer had been correlated with metastasis and poor prognosis [9]. Thus GNT-V could represent a suitable mRNA biomarker for circulating tumor cell detection in gastrointestinal cancer patients.

The present study aims to identify novel GI cancer-specific markers for circulating tumor cells (CTC) detection. The development of mRNA-based tools for CTC biomarkers was done following the proposed guidelines of the Early Detection Research Network (EDRN) by the National Cancer Institute [10]. Phase I preclinical study was performed by means of computational tools for expression analysis. In order to validate the usefulness of this *in silico* approach to identify tumor cells markers, selected genes were evaluated in cell lines and clinical specimens. Finally, qRT-PCR assays for the novel targets plakophilin 3 (PKP3) and anterior gradient-2 (AGR2) to identifiy blood-borne cells in cancer patients were developed.

2. Materials and methods

2.1. Digital differential display and computational tools

Digital differential display (DDD) [11] is an *in silico* data-mining resource of the National Centre for Biotechnology Information (NCBI) for comparing sequence-based gene representation profiles. DDD is based on the UniGene concept and provides the genes that are represented at different levels in the selected cDNA libraries using statistical tests. DDD was used to identify molecular markers highly expressed in GI cancers but absent in blood and bone marrow-derived EST libraries. A set of cDNA libraries from blood and bone marrow ($n = 13$) was selected and it was compared with three pools of colon ($n = 10$), gastric ($n = 13$), and pancreatic ($n = 5$) carcinoma cDNA libraries. Each set was selected to include the broad heterogeneity of every tumor type and includes the minimum possible number of libraries generated from normalized or subtracted cDNAs.

Tags with statistically significant ($p < 0.05$; Fisher's exact test) higher expression in GI cancer libraries were detected and each EST was identified by UniGene cluster.

ESTs whose expression was 10 or more times higher in each set of cancer libraries in comparison with blood and bone marrow-derived EST libraries were selected. As a second filter step, we selected those tags with undetectable or low representation in hematopoietic control libraries, defined as an expression count equal or less than four in DDD output.

Furthermore, the tissue distribution and tumor specificity of every gene selected were analyzed using representation profiles obtained with virtual northern blot [VN, Ref. [12]]. For each combination of tissue and pathologic status, VN shows a spot image representing the expression level of the selected gene and numeric columns. These represent the number of ESTs or SAGE tags corresponding to the gene divided by the total number of ESTs or SAGE tags in all libraries with the given tissue/histology. The statistical significance of the observed counts is measured using Bayesian analysis. Values (p) less or equal to 0.05 were considered significant [13]. Virtual northern blots for EGFR [14] and GNT-V [15] were also obtained. Functional data about the different gene selected were extracted from Gene Ontology (GO) annotations accessed in the Cancer Gene Anatomy Project (CGAP) database [12] ; [16].

2.2. Cell lines

As a surrogated model of gastrointestinal cancer, the following human tumor cell-lines were used: colorectal carcinoma Gp5d, LoVo, DLD1, LS513, HT29, OJC4, OJC5, OJC6; gastric adenocarcinoma OE19, and pancreatic carcinoma OJC1. In addition, hematopoietic cell lines Jurkat, KG1, and K562 were analyzed.

The cell-lines OJC1, OJC4, OJC5, and OJC6 were developed in our laboratory from clinical specimens and they have been fully characterized [17]. All the cell lines were maintained in DMEM (Dulbecco's modified Eagle's medium-high glucose) and MegaCell™ RPMI-1640 mediums (both provided by Sigma–Aldrich) supplemented with 10% fetal calf serum inactivated, 1% penicillin, 1% streptomycin, and 1% amphotericin at 37 °C in 5% CO_2. Cells from adherent cultures were recovered with 1% trypsin–1% EDTA cell dissociating reagent. Cell pellets from suspension cultures were obtained.

2.3. Clinical tissue specimens and blood preparation

Formalin-fixed and paraffin-embedded tissue (FFPE) from colon cancer specimens (n = 16) were obtained from the Pathology Department of the Juan Canalejo University Hospital. Array tissue apparatus, provided by Durviz, was used to obtain tissue cores with 2 mm diameter, selecting an area enriched for tumor tissue. These FFPE cores were used for specific mRNA detection of gene selected using DDD and computational tools.

Blood samples were prospectively collected from an additional cohort of gastrointestinal carcinoma patients (n = 11) and from age and sex matched non-carcinoma donors (n = 10). All the patients included were in the advanced stage, including eight colorectal cancer stage IV, one stage III rectal cancer before any treatment, and two pancreatic adenocarcinomas (stages II and IV). Blood were collected in 10 ml EDTA-containing tubes. To eliminate skin-plug contamination of the blood sample from initial venipuncture, the first several milliliters of blood were discarded. Samples were subjected to lysis and stabilized, in less than 1 h after blood withdrawal, in guanidinium-based RNA/DNA stabilization reagent for blood/bone marrow (Roche, Mannheim, Germany) at 10% (v/v) without cell and plasma separation. The mixtures were stored at −80 °C until mRNA extraction. Isolation reagent for blood and bone marrow (Roche, Mannheim, Germany) was used for mRNA extraction.

The study was approved by the Institutional Review Board of the Ethic Committee of Clinical Investigation of Galicia (Spain) and written informed consent was obtained from all patients.

2.4. RNA extraction and qRT-PCR

Isolation of total RNA from cell cultures was performed using High Pure RNA Isolation Kit (Roche, Mannheim, Germany) following manufacturer's instructions. From each cell line at 70% confluence, 10^6 cells were obtained for RNA isolation. Total RNA was treated with DNase I.

RNA isolation from FFPE tissue cores of each specimen was performed with Optimum™ FFPE RNA Isolation Kit (Ambion, Diagnostics); manufacturer's protocol was followed.

RNA isolated from human normal bone marrows were purchased from BD Biosciences-Clontech. Total RNA was derived from three different pools of normal human bone marrows (n = 23).

mRNA isolation from blood samples was performed with mRNA Isolation Kit for Blood/Bone Marrow (Roche, Mannheim, Germany), following manufacturer's recommendations. Briefly, total nucleic acid fraction was adsorbed to magnetic glass particle and poly(A)+ RNA was captured by using biotin-labeled oligo(dT) and streptavidin-coated magnetic particles. Each mRNA preparation was eluted in 12 μl RNase-free redistilled water. Purified poly (A)+ RNA was further processed in RT-PCR or stored at −80 °C until use. RNA integrity was confirmed by 2% agarose gel electrophoresis and stained with ethidium bromide.

2.5. cDNA synthesis

Reverse-transcription (RT) was performed from total cellular and bone marrow RNA using SuperScript™ First-Strand Synthesis System for RT-PCR (Invitrogen™, USA) up to a total volume of 20 μl × 1 μg of total RNA, 2.5 nM random hexamers, 0.5 mM dNTP mix, and 3 μl of DEPC-treated water were denatured at 65 °C for 5 min and chilled on ice for at least 1 min. On the other hand, 2 μl of 10× RT buffer, 5 mM $MgCl_2$, 0.01 M DTT, and 40 U of RNaseOUT Recombinant Ribonuclease Inhibitor were mixed, collected by centrifugation, and incubated at 25 °C for 2 min. After incubation, 50 U of SuperScript™ RT) were added and incubated at 25 °C

for 10 min, 42 °C for 50 min, and 70 °C for 15 min in a Thermocycler (Gene Amp PCR System 9700, Applied Biosystem). Finally, samples were chilled on ice and incubated with 2 U of RNAse H for 20 min at 37 °C before proceeding to amplification the target cDNA. For mRNA obtained from blood we follow the same procedure and 0.02 μg were subjected to cDNA synthesis. Positive and negative controls were included in each experiment. RNA extraction, reverse transcription-PCR assay setup, and post-reverse transcription-PCR product analysis were carried out in separate designated rooms to prevent cross-contamination. cDNA was quantified and assessed for purity using a GENIOUS UV spectrophotometer. cDNA concentration was measured at 260 nm. Also A260/A280 relation was calculated to know cDNA quality.

2.6. Quantitative real-time reverse transcription-PCR analysis

Real-time PCR analysis was performed, using primers and conditions shown in Table 1, on LightCycler® 480 Instrument (Roche, Mannheim, Germany) using LightCycler 480 SYBR Green I Master (Roche, Mannheim, Germany). PCR reaction consisted of 10 μl of Master Mix 2× conc., 0.35 μM of each forward and reverse primer, template cDNA and PCR-grade water up to a final volume of 20 μl in the LightCycler 480 Multiwell Plate 96. Multiwell Plate was centrifuged at 3000 rpm for 2 min and was loaded in the LightCycler 480 Instrument until the PCR program started. An initial activation at 95 °C for 5 min was followed by an amplification target sequence 50 cycles of 95 °C for 10 s, 54–60 °C (depending on the primers pair used) for 10 s, and 72 °C (depending on the amplicon size amplified) were used. For melting curve analysis 1 cycle of 95 °C for 5 s, 70 °C for 15 s, and 95 °C for 1 s was used. Finally, a cooling step was used at 40 °C for 10 s.

Table 1. qRT-PCR—primers and conditions: primers sequences, conditions and annealing temperature of real-time quantitative RT-PCR used to amplify selected genes

Primers	Sequence	Length (mer)	%GC	Annealing temperature (°C)
PKP3 1R	5′-ggatgaaaggttccacagga-3′	20	50	60
PKP3 2F	5′-ggcccgcagccttcaggccgtgcc-3′	24	79	
GNT-V 1F	5′-att ggc aag cca act ctg a-3′	19	47	55
GNT-V 1R	5′-ttg agg tca aca gtc cac aca-3′	21	48	
AGR2 2F	5′-ctg gcc aga gat acc aca gtc-3′	21	57	54
AGR2 2R	5′-agt tgg tca ccc caa cct-3′	19	58	
S100A16 2F	5′-tgg aga gga ggc aga ctg ag-3′	20	60	54
S100A16 2R	5′-cca cca gga caa tga ctg c-3′	19	58	
LGALS4 1F	5′-aac ctt caa ccc gcc tgt-3′	18	56	54
LGALS4 1R	5′-gag ccc acc ttg aag ttg at-3′	20	50	
LCN2 1F	5′-cag gac tcc acc tca gac ct-3′	20	60	54
LCN2 1R	5′-cac ata cca ctt ccc ctg ga-3′	20	55	
MGC3047 1F	5′-cca gaa gtc ggg aaa gtc aa-3′	20	50	54
MGC3047 1R	5′-ctc act cct gta aag cat ctg g-3′	22	50	
EFNB1 1F	5′-tca tga agg ttg ggc aag a-3′	19	47	54
EFNB1 1R	5′-gtg tgg cca tct tga cag tg-3′	20	55	
ASS 1F	5′-aag ctt ggg gcc aaa aag-3′	18	50	54
ASS1R	5′-gta gcg gtc ctc ata cag tgc-3′	21	57	
CLDN3 2F	5′-cca tta tcc ggg act tct aca ac-3′	23	48	60
CLDN3 2R	5′-gac acg agc agc aga gca-3′	18	61	
EGFR 2F	5′-cag cca ccc ata tgt acc atc-3′	21	52	60
EGFR 2R	5′-aac ttt ggg cga cta tct gc-3′	20	50	
S100A6 2F	5′-act gcg aca cag ccc atc-3′	18	61	60
S100A6 2R	5′-gaa gat ggc cac gag gag-3′	18	61	
HPRT 1F	5′-tga cct tga ttt att ttg cat acc-3′	24	33	60
HPRT 1R	5′-cga gca aga cgt tca gtc ct-3′	20	55	

We verified that amplifications and the expected size of each PCR product were specific. 1.8% agarose gel electrophoresis of all PCR products revealed a single band that corresponded to the single-amplified products as predicted by the melting curve analysis of the PCR. The amplification efficiency was determined for both target and housekeeping genes and was equal (99–100%).

PCR primers for mRNA amplification of the different selected genes were carefully designed using the web-based ProbeFinder software (Universal ProbeLibrary Design Center) [18] or via Roche Applied Science home page [19]. PCR primers have been positioned to span exon–intron boundaries, reducing the risk of detecting genomic DNA. Primers were purchased from Invitrogen™ (USA) and Roche (Mannheim, Germany).

Suitable selection of housekeeping gene(s) was performed using Human Endogenous Control Gene Panel (TATAA Biocenter, Sweden). The Excel macro GeNorm VBA applet for Microsoft Excel was used to determine the gene(s) with most correlated expression in the set of samples. *Homo sapiens* HPRT (Hypoxanthine–guanine phosphoribosyltransferase) were used as internal control for blood, bone marrow, and cell-lines samples. Also housekeeping gene was used to verify integrity of RNA and efficacy of reverse transcription. Any specimen with inadequate HPRT mRNA was excluded from the study.

Data analysis was performed with LigthCycler 480 Relative Quantification software (Roche, Mannheim, Germany). Relative levels of expression were calculated by the $2^{-\Delta\Delta Ct}$ method [20]. Each assay was done at least in triplicate and included marker-positive and marker-negative controls and reagent with no template controls.

2.7. DNA sequencing

At least one PCR product coming from each Real-time PCR experiment was used as template DNA. Products were purified by enzymatic method (ExoSAP-IT, Amersham USB). DNA sequencing was performed in a reference facility on ABI 3700 (Applied Biosystems) using Big Dye terminators. Forward and reverse primers used in sequencing reactions were the same as for the Real-time PCR, see Table 1.

2.8. Statistical analysis

Statistical significance of differences was evaluated at the 95% confidence level by non-parametric statistic, Mann–Whitney *U*-test; p-value <0.05 were considered to be significant. Receiver operating characteristics (ROC) curve analysis was used to evaluate the diagnostic discrimination power of the selected differentially expressed genes. The threshold value for optimal sensitivity and specificity of PKP3 and AGR2 mRNA relative expression levels (R.E.L.) was also determined by ROC analysis. All of the statistical analyses were performed using the SPSS 12.0 for Windows.

3. Results

3.1. Search for gastrointestinal tumor-associated biomarkers using computational tools

Digital differential display (DDD) data were used to identify and prioritize molecular markers highly expressed in GI cancers but absent in blood and bone marrow-derived libraries. The pools used in this analysis, their ID libraries and the ESTs clustered are shown in Table 2. Differential over representation were found for 34 genes in colorectal cancer libraries, for 96 in gastric cancer libraries, and for 144 in pancreatic cancer libraries in comparison with hematopoietic libraries. Potential up-regulated targets of note from this study include, among others, genes involved in cell adhesion and cytoskeleton, cell signaling, growth factor activity, ribosomal proteins, calcium-related metabolism, heat shock proteins as well as hypothetical proteins.

Table 2. Digital differential display: cDNA libraries: Pool of libraries, their identification numbers (IDs) and the expressed sequence tags (ESTs) clustered used in the digital differential display (DDD) bioinformatics tool

Pool	Lib (IDs)	Clustered ESTs
Colorectal cancer	842, 841, 840, 882, 1540, 988, 956, 987, 1447, 486	43918
Gastric cancer	10324, 10311, 10310, 10306, 10325, 10302, 10301, 10299, 10305, 14437, 1449, 721, 733	44581
Pancreatic cancer	1460, 5551, 9885, 721, 733	42014
Blood and bone marrow	7038, 7037, 6976, 6975, 8975, 11923, 931, 5948, 5566, 9724, 8613, 14381, 765	51208

From this set, we specifically selected those ESTs whose expression was at least 10-fold higher in each set of cancer libraries and undetectable or low represented in hematopoietic control libraries, defined as an expression count equal or less than four in DDD output. After these filters, we obtained a collection of 30, 58, and 91 genes in colon, gastric and pancreatic cancer-derived cDNA libraries, respectively (Fig. 1 and Table 3).

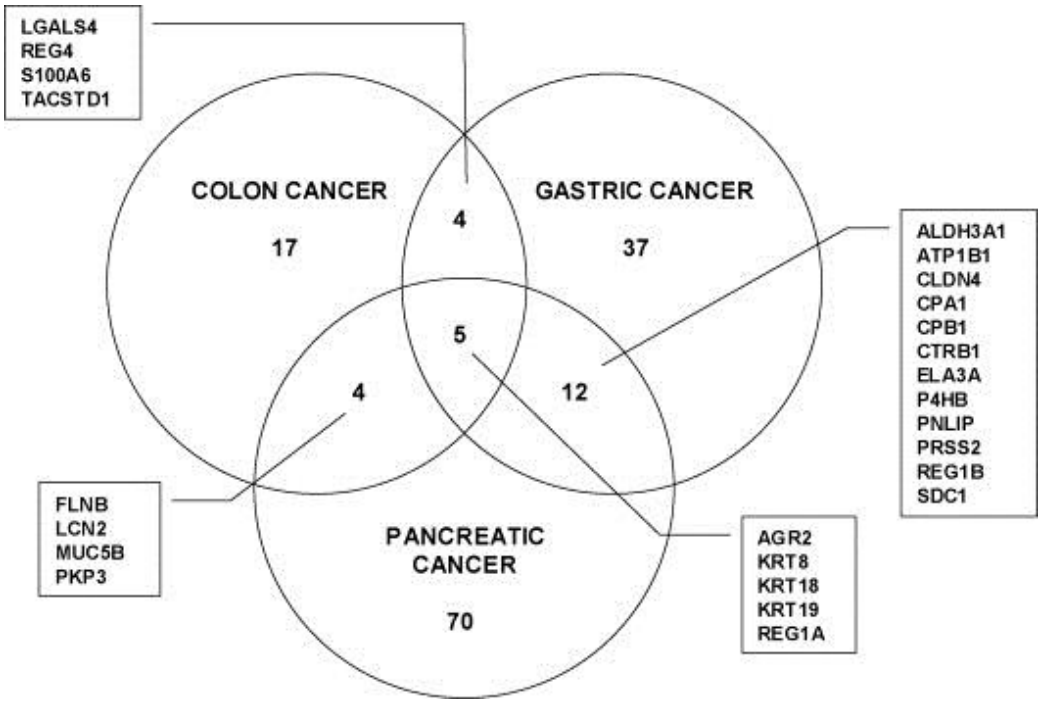

Fig. 1. Overexpressed genes in colon, gastric and pancreatic cancer-derived cDNA libraries: Venn diagram showing overexpressed genes in colon, gastric and pancreatic cancer-derived cDNA libraries. Only those genes whose expression was at least 10-fold higher in each set of cancer libraries and undetectable or low represented in hematopoietic control libraries are shown.

Table 3. Up-regulated genes in gastrointestinal tumors according to digital differential display (DDD)

Gene name	Gene Description	UniGene #	Function	Folds	EST counts in tumor libraries	EST counts in hematopoietic libraries
A: Colorectal						
AGR2	Anterior gradient 2 homolog (*Xenopus laevis*)	Hs.226391	Secretory protein	14	42	3
ASS	Argininosuccinate synthetase	Hs.160786	Enzyme	24	24	0
CEACAM5	Carcinoembryonic antigen-related cell adhesión molecule 5	Hs.220529	Receptor/surface/membrane	22	22	0
CLDN3	Claudin 3	Hs.25640	Receptor/surface/membrane	39	39	0
DIPA	Hepatitis delta antigen-interacting protein A	Hs.66713	Binding protein	20	20	0
EFNB1	Ephrin-B1	Hs.144700	Binding protein	20	20	0
FLNB	Filamin B, beta (actin binding protein 278)	Hs.81008	Binding protein	26	26	1
GDF15	Growth differentiation factor 15	Hs.296638	Secretory protein	41	41	1
GMDS	GDP-mannose 4,6-dehydratase	Hs.105435	Enzyme	22	22	1
H1FX	H1 histone family, member X	Hs.75307	Binding protein	14	56	4
KLF5	Kruppel-like factor 5 (intestinal)	Hs.84728	Binding protein	18	18	0
KRT18	Keratin 18	Hs.406013	Receptor/surface/membrane	42	126	3
KRT19	Keratin 19	Hs.309517	Receptor/surface/membrane	37	37	0
KRT8	Keratin 8	Hs.356123	Receptor/surface/membrane	78	78	1
LCN2	Lipocalin 2 (oncogene 24p3)	Hs.204238	Secretory protein	19	19	0
LGALS4	Lectin, galactoside-binding, soluble, 4 (galectin 4)	Hs.5302	Binding protein	30	30	0
LOC374395	Similar to RIKEN cDNA 1810059G22	Hs.381134	Hypothetical protein	21	21	0
MGC3047	Hypothetical protein	Hs.76239	Hypothetical protein	25	25	1
MUC5B	Mucin 5, subtype B, tracheobronchial	Hs.103707	Secretory protein	24	24	0
PKP3	Plakophilin 3	Hs.148074	Receptor/surface/membrane	28	28	0
REG1A	Regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	Hs.49407	Secretory protein	27	27	1
REG4	Regenerating islet-derived family, member 4	Hs.105484	Secretory protein	25	25	1
RHOB	Ras homolog gene family, member B	Hs.406064	Enzyme	29	29	1
S100A16	S100 calcium binding protein A16	Hs.8182	Binding protein	23	23	0
S100A6	S100 calcium binding protein A6 (calcyclin)	Hs.275243	Binding protein	10	39	4
SAFB	Scaffold attachment factor B	Hs.23978	Binding protein	18	18	0
SLC39A8	Solote carrier family 39 (zinc transporter), member 8	Hs.284205	Zinc transporter	14	28	2
TACSTD1	Tumor-associated calcium signal transducer 1	Hs.692	Receptor/surface/membrane	19	19	0
WARP	Von Willebrand factor A domain-related protein	Hs.449009	Secretory protein	22	22	0
ZFP36L2	Zinc finger protein 36, C3H type-like 2	Hs.78909	Binding protein	12	25	2
B: Gastric						
A2M	Alpha-2-macroglobulin	Hs.74561	Binding protein	37	37	1

AGR2	Anterior gradient 2 homolog (*Xenopus laevis*)	Hs.226391	Secretory protein	35	106	3
ALDH1A1	Aldehyde dehydrogenase 1 family, member A1	Hs.76392	Enzyme	11	44	4
ALDH3A1	Aldehyde dehydrogenase 3 family, member A1	Hs.575	Enzyme	45	45	0
ANKRD9	Ankyrin repeat domain 9	Hs.432945	Binding protein	43	43	0
ATP1B1	ATPase, Na^+/K^+ transporting, beta 1 polypeptide	Hs.78629	Enzyme	25	25	0
CLDN4	Claudin 4	Hs.5372	Receptor/surface/membrane	27	27	1
CPA1	Carboxypeptidase A1 (pancreatic)	Hs.2879	Enzyme	53	53	0
CPB1	Carboxypeptidase B1 (tissue)	Hs.437028	Enzyme	20	20	0
CREB3L1	cAMP responsive element binding protein 3-like 1	Hs.405961	Biniding protein	22	22	0
CTRB1	Chymotrypsinogen B1	Hs.449833	Enzyme	29	29	0
DDX41	DEAD (Asp-Glu-Ala-Asp) box polypeptide 41	Hs.274317	Binding protein	14	59	4
DECR2	2,4-Dienoyl CoA reductase 2, peroxisomal	Hs.15898	Enzyme	19	19	0
ELA3A	Enastase 3A, pancreatic (protease E)	Hs.471034	Enzyme	27	27	0
GLTSCR2	Glioma tumor suppressor candidate region gene 2	Hs.421907	Unknown function	11	47	4
GPX2	Glutathione peroxidase 2 (gastrointestinal)	Hs.2704	Enzyme	21	21	0
GRN	Granulin	Hs.180577	Binding protein	29	29	1
ID1	Inhibitor of DNA binding 1, dominant negative helix-loop-helix protein	Hs.410900	Binding protein	19	19	0
JARID1C	Jumonji, AT-rich interactive domain 1C (RBP2-like)	Hs.103381	Binding protein	27	27	1
KLF13	Kruppel-like factor 13	Hs.7104	Binding protein	18	18	0
KREMEN2	Kringle containing transmembrane protein 2	Hs.351474	Receptor/surface/membrane	25	25	0
KRT18	Keratin 18	Hs.406013	Receptor/surface/membrane	10	32	3
KRT19	Keratin 19	Hs.309517	Receptor/surface/membrane	21	21	0
KRT8	Keratin 8	Hs.356123	Receptor/surface/membrane	146	146	1
LDHA	Lactate dehydrogenase A	Hs.2795	Enzyme	17	383	22
LGALS4	Lectin, galactoside-binding, soluble, 4 (galectin 4)	Hs.5302	Binding protein	19	19	0
LISCH7	Liver-specific bHLH-Zip transcription factor	Hs.312129	Binding protein	22	22	1
LRFN4	Leucine-rich repeat and fibronectin type III domain containing 4	Hs.148438	Binding protein	43	43	1
LYZ	Lysozyme (renal amyloidosis)	Hs.234734	Enzyme	16	32	2
MEIS3	Meis 1, myeloid ecotropic viral integration site 1 homolog 3 (Mouse)	Hs.380923	Binding protein	19	19	0
MUC1	Mucin 1, transmembrane	Hs.89603	Secretory protein	26	26	0
MYEOV	Myeloma overexpressed gene	Hs.436000	Unknown function	21	21	0
MYH11	Myosin, heavy polypeptide 11, smooth muscle	Hs.78344	Binding protein	31	31	0
NQO1	NAD(P)H dehydrogenase, quinone 1 (NQO1)	Hs.406515	Enzyme	10	42	4

P4HB	Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase)	Hs.410578	Binding protein	14	56	4
PGA5	Pepsinogen 5, group I	Hs.432854	Enzyme	122	122	1
PGC	Progastricsin (pepsinogen C)	Hs.1867	Enzyme	95	95	0
PLK2	Polo-like kinase2 (*Drosophila*)	Hs.398157	Enzyme	29	29	1
PME-1	Protein phsphatase methylesterase-1	Hs.63304	Enzyme	15	30	2
PNLIP	Pancreatic lipase	Hs.102876	Enzyme	20	20	0
PRSS2	Protease, serine, 2 (trypsine 2)	Hs.511525	Enzyme	91	91	1
PYCR1	Pyrroline-5-carboxylate reductase 1	Hs.458332	Enzyme	11	71	6
REG1A	Regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	Hs.49407	Secretory protein	55	55	1
REG1B	Regenerating islet-derived 1 beta (pancreatic stone protein, pancreatic thread protein)	Hs.4158	Secretory protein	23	23	0
REG4	Regenerating islet-derived family, member 4	Hs.105484	Secretory protein	68	68	1
RGS5	Regulator of G-protein signalling 5	Hs.24950	Binding protein	30	30	0
S100A6	S100 calcium binding protein A6 (calcyclin)	Hs.275243	Binding protein	11	44	4
SDC1	Syndecan 1	Hs.82109	Binding protein	13	27	2
SERPINB6	Serine (or cysteine)proteinase inhibitor, clade B (ovalbumin), member 6	Hs. 41072	Enzyme	13	27	2
SKB1	SKB1 homolog (*S. pombe*)	Hs.367854	Binding protein	18	18	0
SPARCL1	SPARC-like 1 (mast9, Kevin)	Hs.75445	Secretory protein	20	20	0
TACSTD1	Tumor-associated calcium signal transducer 1	Hs.692	Receptor/surface/membrane	29	29	0
TGFB1I4	Transforming growth factor beta 1-induced transcript 4	Hs.114360	Secretory protein	23	23	1
THBS3	Thrombospondin 3	Hs.169875	Binding protein	20	20	0
TM4SF3	Transmembrane 4 superfamily member 3	Hs.84072	Receptor/surface/membrane	16	32	2
TRY2	Trypsin II precursor	Hs.367767	Enzyme	24	24	0
WBSCR21	Williams Beuren síndrome chromosome region 21	Hs.182476	Binding protein	20	20	0
ZNF499	Zinc finger protein 499	Hs.445346	Binding protein	45	45	0
C: Pancreatic						
ABCD1	ATP-binding cassette, sub-family D (ALD), member 1	Hs.159546	Binding protein	34	34	1
ADAM15	A disintegrin and metalloproteinase domain 15 (metargidin)	Hs.312098	Enzyme	31	31	1
AGR2	Anterior gradient 2 homolog (*Xenopus laevis*)	Hs.226391	Secretory protein	42	127	3
AHSA1	AHA1, activator of heat shock 90 kDa protein ATPase homolog 1 (yeast)	Hs.204041	Enzyme	23	23	1
ALDH3A1	Aldehyde dehydrogenase 3 family, member A1	Hs.575	Enzyme	85	85	0
APG4D	APG4D autophagy 4 homolog D (*S. cerevisiae*)	Hs.512799	Enzyme	52	52	0
ATAD3A	ATPase family, AAA domain containing 3A	Hs.467479	Enzyme	21	21	0

ATP1B1	ATPase, Na^+/K^+ transporting, beta 1 polypeptide	Hs.78629	Enzyme	23	23	0
C1S	Complement component 1, s subcomponent	Hs.458355	Enzyme	23	23	1
CAPN1	Calpain 1 (mu/I) large subunit	Hs.356181	Enzyme	31	31	1
CCT7	Chaperonin containing TCP1, subunit 7 (eta)	Hs.368149	Binding protein	22	89	4
CDC42EP4	CDC42 effector protein (Rho GTPase binding) 4	Hs.3903	Binding protein	180	180	0
CLDN4	Claudin 4	Hs.5372	Receptor/surface/membrane	37	37	1
CPA1	Carboxypeptidase A1 (pancreatic)	Hs.2879	Enzyme	55	55	0
CPB1	Carboxypeptidase B1 (tissue)	Hs.437028	Enzyme	28	28	0
CRABP2	Cellular retinoic acid binding protein 2	Hs.183650	Binding protein	21	21	0
CTRB1	Chymotrypsinogen B1	Hs.74502	Enzyme	18	18	0
CTSD	Cathepsin D (lysosomal aspartyl protease)	Hs.34375	Enzyme	10	31	3
DHRSX	Dehydrogenase/reductase (SDR family) X-linked	Hs.11779	Enzyme	53	53	0
DKK3	Dickkopf homolog 3 (*Xenopus laevis*)	Hs.130865	Secretory protein	99	99	0
DLGAP4	Discs, large (*Drosophila*) homolog-associated protein 4	Hs.249600	Binding protein	18	36	2
DRAP1	DR1-associated protein 1 (negative cofactor 2 α)	Hs.356742	Binding protein	24	24	1
EEF1A2	Eukaryotic translation elongation factor 1 alpha 2	Hs.433839	Binding protein	59	59	0
EGLN2	Egl nine homolog 2 (*C. elegans*)	Hs.324277	Binding protein	29	29	0
EIF4G1	Eukaryotic translation initiation factor 4 gamma, 1	Hs.433750	Binding protein	13	26	2
ELA3A	Enastase 3A, pancreatic (protease E)	Hs.471034	Enzyme	33	33	0
ELF3	E74-like factor 3 (ets domain transcription factor, ephitelial specific)	Hs.67928	Binding protein	20	20	1
EPHA2	EphA2	Hs.171596	Receptor/surface/membrane	19	19	0
FKBP4	FK506 binding protein 4, 59 kDa	Hs.848	Binding protein	31	63	2
FLJ23322	Hypothetical protein	Hs.387601	Hypothetical protein	58	58	1
FLNB	Filamin B, beta (actin binding protein 278)	Hs.81008	Binding protein	30	30	1
FLNC	Filamin C, gamma (actin binding protein 280)	Hs.58414	Binding protein	67	67	0
GCG	Glucagon	Hs.423901	Enzyme	74	74	0
GSN	Gelsolin (amyloidosis, Finnish type)	Hs.446537	Binding protein	10	30	3
HAS3	Hyaluronan synthase 3	Hs.85962	Enzyme	23	23	1
HMGA1	High mobility group AT-hook 1	Hs.57301	Binding protein	21	108	5
IDS	Iduronate 2-sulfatase (Hunter syndrome)	Hs.303154	Enzyme	21	21	0
IFI27	Interferon, alpha-inducible protein 27	Hs.278613	Receptor/surface/membrane	25	25	1
IGFBP6	Insulin-like growth factor binding protein 6	Hs.274313	Binding protein	22	22	0
ITGB4	Integrin, beta 4	Hs.85266	Cell adhesion molecules	19	19	0
ITGB6	Integrin, beta 6	Hs.57664	Cell adhesion molecules	23	23	0
JUP	Junction plakoglobin	Hs.2340	Cell adhesion molecules	57	57	1
KRT13	Keratin 13	Hs.433871	Receptor/surface/membrane	28	28	0

KRT17	Keratin 17	Hs.2785	Receptor/surface/membrane	37	112	3
KRT18	Keratin 18	Hs.406013	Receptor/surface/membrane	116	349	3
KRT19	Keratin 19	Hs.309517	Receptor/surface/membrane	163	163	0
KRT5	Keratin 5	Hs.433845	Receptor/surface/membrane	79	79	0
KRT6A	Keratin 6A	Hs.367762	Receptor/surface/membrane	37	37	0
KRT7	Keratin 7	Hs.23881	Receptor/surface/membrane	39	79	2
KRT8	Keratin 8	Hs.356123	Receptor/surface/membrane	223	223	1
KRTHB1	Keratin, hair, Basic, 1	Hs.170925	Receptor/surface/membrane	34	34	1
LAMC2	Laminin, gamma 2	Hs.54451	Cell adhesion molecule	27	27	0
LCN2	Lipocalin 2 (encogen 24p3)	Hs.204238	Secretory protein	55	55	0
LOC142678	Skeletrophin	Hs.135805	Binding protein	41	41	1
LRP1	Low density lipoprotein-related protein 1(alpha-2-macroglobulin receptor)	Hs.162757	Receptor/surface/membrane	13	27	2
LUM	Lumican	Hs.406475	Binding protein	10	63	6
LY6E	Lymphocyte antigen 6 complex, locus E	Hs.77667	Receptor/surface/membrane	21	42	2
MGAT4B	Mannosyl (alpha-1,3)-glycoprotein beta-1,4-*N*-acetylglucosaminyltransferase, isoenzyme B	Hs.437277	Enzyme	69	69	0
MGC10471	Hypothetical protein	Hs.24998	Hypothetical protein	109	109	1
MGC2776	Hypothetical protein	Hs.335550	Hypothetical protein	49	49	0
MGLL	Monoglyceride lipase	Hs.409826	Enzyme	21	21	1
MRPL49	Mitochondrial ribosomal protein L-49	Hs.75859	Ribosomal protein	11	59	5
MRPS12	Mitochondrial ribosomal protein S12	Hs.411125	Ribosomal protein	23	23	0
MUC5B	Mucin 5, subtype B, tracheobronchial	Hs.103707	Secretory protein	34	34	0
NID	Nidogen (enactin)	Hs.356624	Binding protein	82	82	1
NRBP	Nuclear receptor binding protein	Hs.272736	Binding protein	16	50	3
NUMBL	Numb homolog (*Drosophila*)-like	Hs.326953	Unknown function	20	20	0
P4HB	Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase)	Hs.410578	Enzyme	14	57	4
PAK4	p21 (CDKN!A)-activated kinase 4	Hs.20447	Binding protein	26	26	0
PCBD	6-Pyruvoyl-tetrahydropterin synthase/dimerization cofactor of hepatocyte nuclear factor 1 alpha (TCF1)	Hs.3192	Enzyme	33	33	1
PCQAP	PC2 (positive cofactor 2, multiprotein complex) glutamina/Q-rich-associated protein	Hs.410347	Transcription regulatory protein	21	43	2
PKP3	Plakophilin 3	Hs.148074	Receptor/surface/membrane	23	23	0
PLAT	Plasminogen activator, tissue	Hs.274404	Enzyme	32	32	1
PLEC1	Plactin 1, intermediate filament binding protein 500 kDa	Hs.79706	Binding protein	40	40	1
PNLIP	Pancreatic lipase	Hs.102876	Enzyme	24	24	0
PRSS2	Protease, serine, 2 (trypsin 2)	Hs.511525	Enzyme	94	94	1
PRSS3	Protease, serine, 3 (mesotrypsin)	Hs.435699	Enzyme	31	31	1
RAC1	Ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	Hs.413812	Binding protein	18	73	4

REG1A	Regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	Hs.49407	Secretory protein	46	46	1
REG1B	Regenerating islet-derived 1 beta (pancreatic stone protein, pancreatic thread protein)	Hs.4158	Secretory protein	24	24	0
RGS19IP1	Regulador of G-protein signalling 19 interacting protein 1	Hs.6454	Receptor/surface/membrane	12	24	2
SCD	Stearoyl-CoA desaturase (delta-9-desaturase)	Hs.119597	Enzyme	29	59	2
SCNN1A	Sodium channel, nonvoltage-gated 1 alpha	Hs.446415	Binding protein	22	22	0
SDC1	Syndecan 1	Hs.82109	Receptor/surface/membrane	12	24	2
SERPINA1	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	Hs.297681	Enzyme	26	26	0
SFRP2	Secreted frizzled-related protein 2	Hs.31386	Secretory protein	29	29	0
SSFA2	Sperm specific antigen 2	Hs.438599	Receptor/surface/membrane	41	41	0
STMN3	Stathmin-like 3	Hs.285753	Binding protein	12	63	5
SUGT1	SGT1, supressor of G2 allele of SKP1 (S. cerevisiae)	Hs.107776	Binding protein	24	24	1
TIMM50	Translocase of inner mitochondrial membran 50 homolog (yeast)	Hs.355819	Enzyme	17	51	3
TMEPAI	Transmembrane, prostate androgen-induced RNA	Hs.83883	Unknown function	20	40	2

Hits (known ESTs) showing ≥10-fold differences in the tumor-derived libraries in comparison with normal blood and bone marrow tissue-derived cDNA libraries are shown: colorectal (A), gastric (B) and pancreatic (C). Gene name, gene description, gene function and UniGene number of each hit are also shown. EST counts in tumors and hematopoietics libraries are shown. The last column of the tables indicates the number of fold differences in the gastrointestinal tumor-derived libraries in comparison with normal tissue-derived cDNA libraries. When EST counts in hematopoietics libraries were zero, EST counts in tumors are considered as fold difference.

Among all of the up-regulated genes obtained, we identified several known genes which have previously been shown to be overexpressed in gastrointestinal tumors, and proposed as micrometastasis markers [21] as keratins 8, 18, and 19, CEACAM5, and tumor-associated calcium signal transducer 1 (TACSTD1). The tissue distribution and tumor specificity were analyzed using representation profiles obtained with virtual northern (Table 4). Based on *in silico* data we could predict that some of the markers could be expressed in the hematopoietic compartment.

Table 4. Expression of selected markers in gastrointestinal cancers, bone marrow and white blood cells: EST and SAGE data

	Colon cancer		Gastric cancer		Pancreatic cancer		Bone marrow		Blood
	EST Data	SAGE data	EST data	SAGE data	EST data	SAGE data	EST data	SAGE data	SAGE data
PKP3	35/170084	27/643586	11/70532	74/448716	23/75487	13/189999	0/44611	0/282890	4/846268
AGR2	90/151938	45/643586	117/65595	342/448716	133/74783	36/189999	3/44571	0/282890	0/846268
LCN2	48/170084	53/643586	17/70532	165/448716	60/75487	213/189999	2/44611	90/282890	2/846268
EFNB1	30/170084	21/643586	1/70532	28/448716	9/75487	7/189999	0/44611	2/282890	11/846268
ASS1	62/170084	147/643586	17/70532	41/448716	20/75487	16/189999	0/44611	0/282890	3/846268
LGALS4	93/170084	56/643586	18/70532	232/448716	8/75487	46/189999	0/44611	0/282890	1/846268
MGC3047	1/170084	0/643586	2/70532	9/448716	6/75487	8/189999	0/44611	0/282890	0/846268
CLDN3	152/170084	80/643586	6/70532	96/448716	1/75487	29/189999	0/44611	0/282890	7/846268
S100A16	46/170084	164/643586	5/70532	73/448716	20/75487	66/189999	2/44611	0/282890	1/846268
S100A6	65/151938	506/643586	49/65595	656/448716	68/74783	480/189999	3/44571	77/282890	625/846268
GNT-V	4/151938	2/643586	1/65595	0/448716	0/74783	0/189999	0/44571	0/282890	0/846268
EGFR	7/151938	1/643586	3/65595	2/448716	0/74783	0/189999	0/44571	0/282890	0/846268

Each column represents the number of ESTs or SAGE tags corresponding to the gene divided by the total number of ESTs or SAGE tags in all libraries with the given tissue/histology. Only SAGE data are available for white cells blood. PKP3, plakophilin 3; AGR2, anterior gradient homolog 2; LCN2, lipocalin 2 (oncogene 24p3); EFNB1, ephrin-B1; ASS1, argininosuccinate synthetase; LGALS4; lectin, galactoside-binding, soluble, 4 (galectin 4); MGC3047, hypothetical protein, MXRA8; CLDN3, claudin 3; S100A16, S100 calcium binding protein A16; S100A6, S100 calcium binding protein A6 (calcyclin); GNT-V, beta-1,6-*N*-acetylglucosaminyltransferase V, MGAT 5; EGFR, epidermal growth factor receptor.

3.2. Experimental validation of the in silico data: qRT-PCR in cell lines

Criteria to select molecular markers identified by bioinformatic approach for quantitative real-time PCR studies are described in "Section 2.1". In brief, we select those novel markers with absent or low expression in blood and bone marrow-derived EST libraries and high expression in gastrointestinal cancer-derived EST libraries. In addition functional annotations based on Gene Ontology and related with epithelial cellular components or (potentially) with metastatic process [22] were taken into account.

In this way, we chose PKP3, AGR2, LCN2, S100A16, S100A6, EFNB1, ASS1, LGALS4, MGC3047/MXRA8, and CLDN3 for further experimental validation.

Interestingly PKP3, LGALS4, CLDN3, and EFNB1 were involved in cell adhesion, following biological process annotations in Gene Ontology. S100A6 and S100A16 are related in calcium-dependent protein binding. ASS1 has argininosuccinate synthase activity. No significantly overrepresented GO annotations were found for AGR2, LCN2 and MGC3047/MXRA8.

The expression of these selected markers was compared against EGFR and GNT-V, both described as overexpressed in gastrointestinal cancer. The expression levels of these genes were measured in 10 gastrointestinal tumor cell-lines (eight colorectal tumor cell-lines, one gastroesophageal cell-line, and one pancreatic cell-line). The values obtained were compared against three hematopoietic tumor cell-lines (Jurkat, K562, and KG1) and three different pools of normal bone marrow that include 23 individual samples. All selected genes assessed showed higher expression levels in gastrointestinal tumor cell-lines than in normal human bone marrow with two exceptions, LCN2 and MGC3047 (Fig. 2). LCN2 showed lower expression levels in all

gastrointestinal tumor cell-lines tested than normal human bone marrow. MGC3047 only showed higher expression levels over to those obtained with normal human bone marrow in OJC1 pancreatic tumor cell-line. The rest of the markers showed high expression in at least 7–10 gastrointestinal tumor cell-lines than in normal human bone marrow. Also, S100A16 was the unique marker assessed that showed higher expression levels than normal human bone marrow in all tumor cell-lines tested. Although GNT-V was expressed in 8 out of 10 cell digestive cancer cell-lines, its values were together with LCN2, EFNB1, ASS1, and MGC3047, the ones that showed the lowest expression in the tumor cell-lines assessed and they were not selected for blood mRNA quantification.

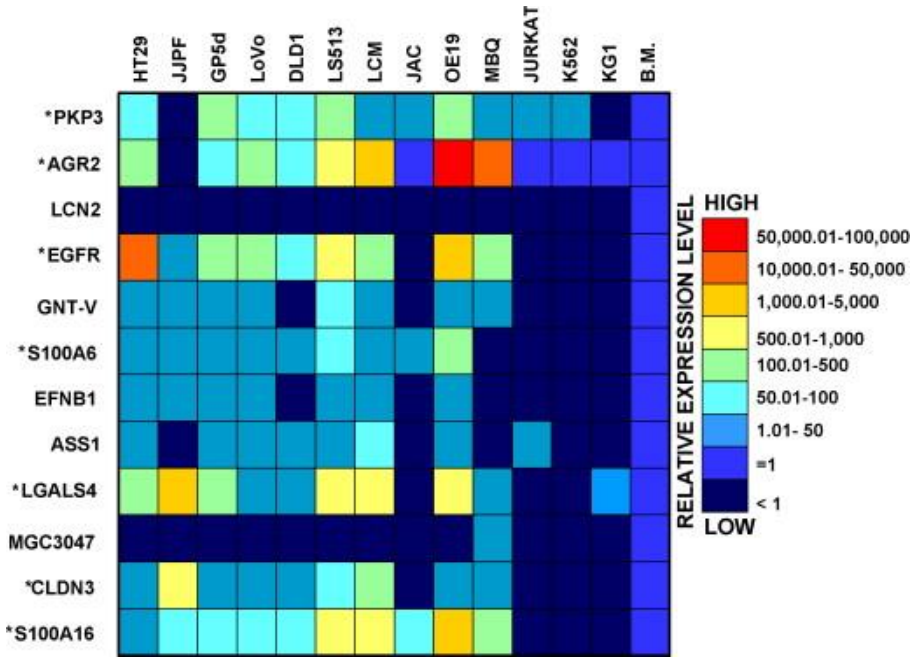

Fig. 2. Expression profile in cell-lines and normal human bone marrows: heat-map showing relative mRNA expression of *in silico* selected genes (rows) in cell-lines and bone marrow (columns). mRNA relative expression were measured by qRT-PCR. Results are the mean of at least three independent measurements. Data were normalized *vs.* the values obtained in human normal bone marrow. Relative expression levels are shown in a color scale. Genes are described in Table 3. Cell-lines (columns) are as shown: colorectal cancer (Gp5d, LoVo, DLD1, LS513, HT29, OJC4, OJC5, OJC6); gastric adenocarcinoma (OE19); pancreatic carcinoma (OJC1); hematopoietic (Jurkat, KG1 and K562); normal human bone marrows (BM). *Indicated those genes selected for their mRNA quantification in blood samples from patients.

Among all the genes assessed, PKP3, AGR2, S100A16, S100A6, LGALS4, and CLDN3 showed the most highly expression values with regard to normal human bone marrow.

3.3. mRNA quantitative detection of selected genes in peripheral blood from patients with gastrointestinal cancer and controls

We chose PKP3, AGR2, S100A6, S100A16, LGALS4, and CLDN3 for mRNA quantitative detection in peripheral blood samples from gastrointestinal cancer patients and controls. All of them showed an expression up to 10^2- to 10^5-fold higher in tumor cell-lines than in normal human bone marrow. Also, EGFR mRNA expression was measured in blood samples in order to compare their results with those obtained with the mRNA markers selected. All of these were expressed in at least 8 out of 10 tumor cell-lines.

Among the six potential molecular markers assessed in-patient and matched control blood samples, PKP3 and AGR2 showed the most remarkable results (Table 5 and Fig. 3). Values for PKP3 and AGR2 mRNA were significantly higher in blood samples from gastrointestinal cancer patients than in blood samples from non-carcinoma patients with a significant level of $p = 0.000$ and 0.019, respectively (Mann–Whitney *U*-test).

Table 5. Median relative expression levels and ratios of candidate mRNA markers of circulating tumor cells in patients and controls blood

Gene name	Gene symbol	Patients blood	Controls blood	Patients: controls
Plakophilin 3	*PKP3*	175.42	9.32	18.82
Anterior gradient 2	*AGR2*	13.50	0	*
Claudin 3	*CLDN3*	8040.27	29,299	2.74
Galectin 4	*LGALS4*	27,073	13,070	2.07
S100 calcium binding protein A6 (calcyclin)	*S100A6*	53,840	58548.6	0.92
S100 calcium binding protein A16	*S100A16*	41,945	23,090	1.82
Epidermal growth factor receptor	*EGFR*	67.22	24.64	2.72

mRNA levels were measured by qRT-PCR as described in Section 2. Results are the mean of at least three independent measurements. *There was no AGR2 expression in normal bloods.

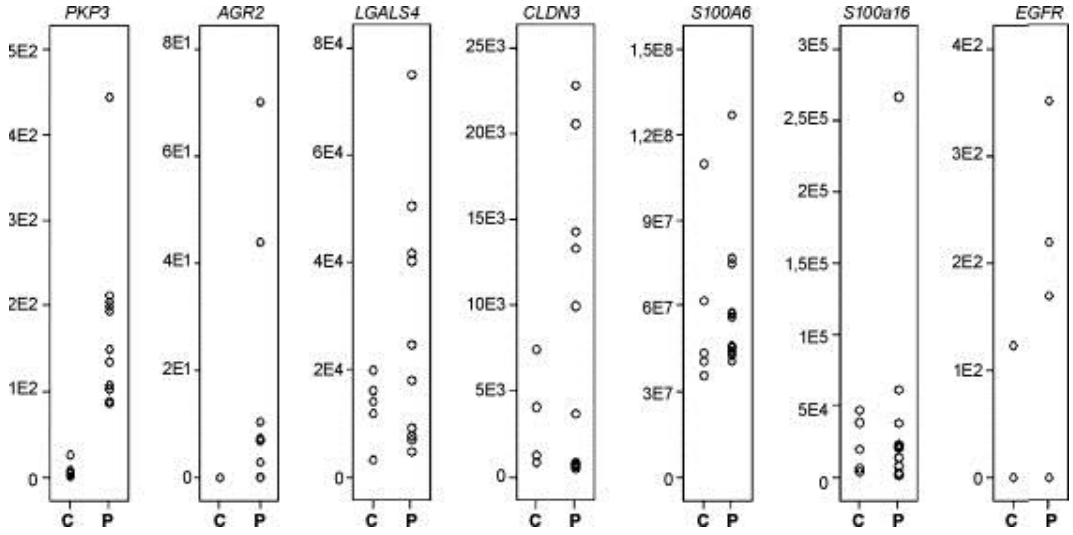

Fig. 3. mRNA expression levels of PKP3, AGR2, LGALS4, CLDN3, S100A6, S100A16 and EGFR mRNA in peripheral blood: mRNA expression levels of *PKP3*, *AGR2*, *LGALS4*, *CLDN3*, *S100A6*, *S100A16* and *EGFR* mRNA (*Y*-axis) determined in peripheral blood from non-carcinoma patients (C) and patients (P) with gastrointestinal cancer (*X*-axis). mRNA levels were measured by qRT-PCR as described in Section 2. Results are the mean of at least three independent measurements.

ROC curve analysis showed that the area under the ROC curve for PKP3 mRNA relative expression levels was 1.00 (S.E. = 0.00; $p = 0.002$). Optimal sensitivity (100%) and specificity (100%) was obtained by applying 56.61 as a PKP3 mRNA R.E.L. cutoff value.

In the case of AGR2, ROC curve analysis showed that the area under the ROC curve for AGR2 mRNA relative expression levels was 0.864 (S.E. = 0.092; 95% confidence range 0.683–1.045; $p = 0.023$). Optimal sensitivity (72.7%) and specificity (100%) was obtained by applying 0.044 as AGR2 mRNA R.E.L. cutoff value. In both cases, the calculated cutoff value may be useful to define positive or negative results and it discriminates samples of peripheral blood from patients with gastrointestinal cancer and control donors.

Area under the ROC curve for EGFR mRNA relative expression levels was 0.564 (S.E. = 0.151; 95% confidence range 0.268–0.859; $p = 0.692$). Thus, in our series, EGFR mRNA amplification was not useful to differentiate between carcinoma patients and controls. In relation to the other four potential molecular markers assessed, S100A6, S100A16, LGALS4, and CLDN3, neither of them could be considered specific enough to detect circulating tumor cells. Perhaps increasing the number of samples analyzed and ROC curve analysis, or using logistic regression to determine a combination equation using all six molecular markers assessed could yield a stronger and powerful diagnostic discrimination test.

PKP3 and AGR2 seem to be two new promising molecular markers for qRT-PCR detection of circulating tumor cells in peripheral blood samples from gastrointestinal cancer patients. Our findings had improved the previous results obtained with other common markers, widely used in micrometastasis detection, such as EGFR.

3.4. mRNA detection of selected genes in colon cancer samples

To validate the DDD data and the results obtained in cell-lines, we further investigated the mRNA expression of the genes PKP3, AGR2, LGALS4, and CLDN3 in colon cancer samples. These genes exhibited up-regulated expression based on both *in silico* data and GI cancer cell-lines qRT-PCR results. Absolute quantification by real-time RT-PCR analysis using RNA isolated from FFPE was used. Patient samples included primary colon tumors ($n = 14$). In addition metachronous metastatic lesions (hepatic and peritoneal) excised from two of these patients were studied.

Among all of the markers tested, CLDN3 showed expression in a large number of tissues analyzed (90.9%) followed by AGR2 (81.8%), LGALS4 (45.4%), and PKP3 (40%). For PKP3, there was no concordant mRNA expression between primary and metastasis in the cases studied, suggesting a dynamic regulation of their expression during tumor progression.

4. Discussion

Metastatic hematogenous spreading is one of the most important factors affecting the prognosis of carcinoma patients, including gastrointestinal cancer. Detection of carcinoma cells in the blood or minimal deposits in distant organs as bone marrow could be important to identify patients at high risk of relapse or disease progression [7]. PCR amplification of tissue or tumor-selective mRNA is the most powerful tool for surrogate detection of this isolated tumor cells. However, specificity of different mRNA-based assays has been controversial and there is a need for a systematic evaluation for each new biomarker proposed.

Our research aimed to identify, at first, a set of specific epithelial gene markers highly expressed in gastrointestinal cancer. Electronic databases analysis allowed us to obtain a list of genes that are differentially expressed, to the point of statistical significance, in colon, pancreas, and gastric libraries. We hypothesized that these transcripts could represent the molecular signature associated with CTC. Global gene expression profile of cancer cells isolated from blood has been recently described in colon, breast, and prostatic tumors, using an immunomagnetic enrichment and microarray analysis [23]. In a recent report, Solmi et al. [24] have used a microarray-based high-throughput screening method to identify candidate marker mRNAs for CTC detection. However, employing a conventional RT-PCR procedure they could not find a set of suitable and consistent mRNA marker for detection of colonic cells in the blood. In previous reports using both *in silico* data-mining tools and conventional RT-PCR, these authors [25] suggest that among 15 genes differentially expressed in colon, only non-SMC structural maintenance of chromosomes, element 1 protein (NSE1) and gastrin (GAS) mRNAs could be suitable for detection of circulating colon cancer cells in blood.

To point out the validity of our results using an *in silico* approach, our set of overexpressed genes included different markers previously used for CTC detection. In addition, AGR2 and different S100 proteins were identified in our work as well as in the CTC molecular profile proposed by Smirnov et al. [23].

To obtain independent evidence to support the bioinformatics-approach for selecting CTC markers, quantitative mRNA expression of selected genes were analyzed in gastrointestinal cell-lines, hematopoietic cell-lines, and normal bone marrows. In order to validate their usefulness as potential molecular markers for gastrointestinal disseminated tumor cell detection, these results were compared with those obtained for EGFR and GNT-V.

Among all the genes assessed, PKP3, AGR2, S100A16, S100A6, LGALS4, and CLDN3 showed the most highly expressed values with regard to normal human bone marrow. For this reason they were chosen for further analysis for their expression in blood from gastrointestinal

cancer patients and controls. One of the main problems with any mRNA-based assays is the illegitimate transcription of the so-called "epithelial specific transcripts" in hematopoietic compartment [2]; [3] ; [4]. In our study, we included a quantitative RT-PCR approach and a significant number of controls samples (23 bone marrows and 10 bloods) in order to select the most specific mRNA markers.

In patient blood samples, qRT-PCR for PKP3 and AGR2 showed the most promising results. We could calculate an mRNA R.E.L. cutoff value of 56.61 and 0.044 for PKP3 and AGR2, respectively. These cut-off points may be useful to distinguish between samples of peripheral blood from gastrointestinal cancer patients and control samples. However, we need to underscore that all of the patients included in our series of blood samples were in advanced disease status. To better evaluate the predictive and prognostic role (if any) of PKP3 and AGR2 qRT-PCR assays in blood, a prospective study with a larger number, and less advanced, cancer patients are clearly needed.

When we compared our results obtained for PKP3 and AGR2 with those obtained for EGFR, our findings suggest that PKP3 and AGR2 could be better for its use as surrogate markers for circulating tumor cells detection in patients with GI cancer.

This is the first time that PKP3, a member of the p120ctn/plakophilin subfamily of armadillo proteins, is suggested as a molecular marker for CTC detection. PKP3 is present in desmosomes of epithelial cells and binds all three desmogleins, desmocollin, plakoglobin, desmoplakin, and the epithelial keratin 18. This protein not only functions in desmosome-dependent cell adhesion, but in signaling transduction [26]; [27] ; [28].

In view of the importance of β-catenin and other armadillo proteins in tumor biology, it is surprising that up to now, knowledge about the occurrence of PKPs in tumors is limited and little is known about their possible biologic role in tumor invasion and metastasis [29]. Recent studies had shown evidences that PKP3 protein may serve as useful marker for predicting the clinical outcome of head and neck tumors [30] and non-small cell lung cancer [31]. Other investigators [32] had demonstrated that this protein is highly expressed in various types of adenocarcinomas (colorectal, pancreatic, and prostate). In our work, PKP3 mRNA was highly expressed in digestive cancer cell-lines (8/10) but less in colon cancer samples (40%). One limitation of our analysis was the use of RNA isolated from FFPE. Recently, we studied PKP3 protein by immunohistochemistry in a large set of digestive tumors and found its expression in 66.7% of cases [33]. Aigner et al. have shown that PKP3 is repressed in the epithelial to mesenchymal transition of tumors under control by the E-cadherin repressor ZEB1 [34]. These data suggest, as previously stated [23] that gene expression signatures of CTCs may change as the disease progresses. In addition, selective suppression of gene transcription was one of the characteristics of the molecular signature associated with bone marrow micrometastasis in breast cancer [35].

In relation with AGR2, the human homolog of Xenopus anterior gradient 2 (XAG2), our results corroborated and complemented a previous study in which it has been postulated as a marker for detection of circulating tumor cells in peripheral blood of advanced cancer patients [23]. AGR2 mRNA is expressed in several normal human tissue types that are rich in epithelial cells, including colon, pancreas, lung, stomach, rectum, prostate, and breast [36]; [37] ; [38]. AGR2 has been demonstrated to be up-regulated not only in hormonal-dependent tumors as breast, endometrial, renal, and prostate cancers but also in lung carcinomas [39]. Using qRT-PCR analysis, we demonstrated an increased AGR2 expression in gastrointestinal cancer cell-lines and colon tumors. Although AGR2 function is not known, experimental data suggest that it could act as a survival factor through inhibition of p53, and enhancing tumor cell adhesion to substratum [40] ; [41].

However, identification of truly specific markers is a difficult task. Our quantitative RT-PCR assay demonstrated background expression in blood derived from non-tumor donors for S100A16, S100A6, LGALS4, and CLDN3. Although these gene transcripts were found highly expressed in cancer specimens, none of them could be considered specific enough, at first, for circulating epithelial cells detection. Their value in multigene qRT-PCR approach as a discrimination test could be analyzed using logistic regression and ROC curve.

In conclusion, novel and known gastrointestinal-specific mRNA markers for circulating tumor cells have been identified through *in silico* analysis and validated in clinical material. Our results indicate that quantitative real-time reverse transcription-PCR assay targeted to PKP3 and AGR2 mRNAs might be helpful to detect circulating tumor cells in patients with gastrointestinal cancer.

Conflict of interest
None declared.

Acknowledgements
This study was supported by grant 5090252501 from Universidade da Coruña. S. Díaz-Prado is supported by an Isidro Parga Pondal research contract by Xunta de Galicia (A Coruña, Galicia, Spain). Cancer research in our laboratory is supported by the "Fundación Juan Canalejo-Marítimo de Oza". We thank the CGAP database for providing access and the data-mining tools used in this study.

References

[1]. M.P. Coleman, G. Gatta, A. Verdecchia, J. Esteve, M. Sant, H. Storm, *et al*. EUROCARE-3 summary: cancer survival in Europe at the end of the 20th century. Ann Oncol, 14 (5 Suppl.) (2003), pp. v128–v149.
[2]. F.A. Vlems, J.H.S. Diepstra, I.M.H.A. Cornelissen, T.J. Ruers, M.J. Ligtenberg, C.J. Punt, *et al.*. Limitations of cytokeratin 20 RT-PCR to detect disseminated tumour cells in blood and bone marrow of patients with colorectal cancer: expression in controls and downregulation in tumour tissue. Mol Pathol, 55 (2002), pp. 156–163.
[3]. N. Dandachi, M. Balic, S. Stanzer, M. Halm, M. Resel, T.A. Hinterleitner, *et al*. Critical evaluation of real-time reverse transcriptase-polymerase chain reaction for the quantitative detection of cytokeratin 20 mRNA in colorectal cancer patients. J Mol Diagn, 7 (2005), pp. 631–637.
[4]. R. Schuster, N. Max, B. Mann, K. Heufelder, F. Thilo, J. Gröne, *et al*. Quantitative real-time RT-PCR for detection of disseminated tumor cells in peripheral blood of patients with colorectal cancer using different mRNA markers. Int J Cancer, 108 (2004), pp. 219–227.
[5]. S.A. Bustin, R. Mueller. Real-time reverse transcription PCR (qRT-PCR) and its potential use in clinical diagnosis. Clin Sci (Lond), 109 (2005), pp. 365–379.
[6]. D.S. Salomon, R. Brandt, F. Ciardiello, N. Normanno. Epidermal growth factor-related peptides and their receptors in human malignancies. Crit Rev Oncol Hematol, 19 (1995), pp. 183–232.
[7]. K. Pantel, C. Alix-Panabieres. The clinical significance of circulating tumor cells. Nat Clin Pract Oncol, 4 (2007), pp. 62–63.
[8]. Y. Sugita, Y. Fujiwara, D.S. Hoon, A. Miyamoto, M. Sakon, C.T. Kuo, *et al.* Overexpression of beta 1,4-*N*-acetylgalactosaminyl-transferase mRNA as a molecular marker for various types of cancers. Oncology, 62 (2002), pp. 149–156.
[9]. K. Murata, E. Miyoshi, M. Kameyama, O. Ishikawa, T. Kabuto, Y. Sasaki, *et al*. Expression of *N*-acetylglucosaminyl transferase V in colorectal cancer correlates with metastasis and poor prognosis. Clin Cancer Res, 6 (2000), pp. 1772–1777.
[10]. M.S. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, *et al*. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst, 93 (2001), pp. 1054–1061.
[11]. UniGene [database on the Internet]. Bethesda: National Library of Medicine (US); [cited Jan 22, 2006]. Digital Differential Display (DDD). *Homo sapiens*. Available from http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs.
[12]. Cancer Genome Anatomy Project [database on the Internet]. Bethesda (MD): National Cancer Institute (US); [cited December 15, 2005]. GeneFinder. Available from http://cgap.nci.nih.gov/Genes/GeneFinder.

[13]. A. Lal, A.E. Lash, S.F. Altschul, V. Vekulescu, L. Zhang, R.E. McLendon, *et al*. A public database for gene expression in human cancers. Cancer Res, 59 (1999), pp. 5403–5407.
[14]. Cancer Genome Anatomy Project [database on the Internet]. Bethesda (MD): National Cancer Institute (US); [cited December 15, 2005]. Tissues. Virtual Northern. EGFR. Available from http://cgap.nci.nih.gov/Tissues/VirtualNorthern?TEXT=0&ORG=Hs&CID=488293.
[15]. Cancer Genome Anatomy Project [database on the Internet]. Bethesda (MD): National Cancer Institute (US); [cited December 15, 2005]. Tissues. Virtual Northern. MGAT5. Available from http://cgap.nci.nih.gov/Tissues/VirtualNorthern?TEXT=0&ORG=Hs&CID=651869.
[16]. Cancer Genome Anatomy Project [database on the Internet]. Bethesda (MD): National Cancer Institute (US); [cited December 15, 2005]. GoBrowser. Available from http://cgap.nci.nih.gov/Genes/GOBrowser.
[17]. M. Valladares Ayerbes, L. Calvo, G. Alonso, P. Iglesias, M.J. Lorenzo, I. Brandón, *et al*. Evaluation of messenger RNA of pituitary tumour-transforming gene-1 (PTTG1) as a molecular marker for micrometastasis. J.J. Li, S.A. Li, A. Llombart-Bosch (Eds.), Hormonal carcinogenesis IV, Springer-Verlag, New York (2005), pp. 462–468.
[18]. Roche Applied Science [database on the Internet] [cited December 20, 2005]. Assay Design Center/ProbeFinder. *Homo sapiens* (Human). Available from http://www.universalprobelibrary.com.
[19]. Roche Applied Science [database on the Internet] [cited December 20, 2005]. Universal ProbeLibrary. Universal ProbeLibrary interest site. Assay Design Center/ProbeFinder. *Homo sapiens* (Human). Available from http://www.roche-applied-science.com.
[20]. K.J. Livak, T.D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method. Methods, 25 (2001), pp. 402–408.
[21]. I. Vogel, H. Kalthoff. Clinical relevance of tumor cell dissemination in colorectal, gastric and pancreatic carcinoma. K. Pantel (Ed.), Micrometastasis, Kluwer Academic Publishers, Dordrecht (2003), pp. 139–172.
[22]. G.P. Gupta, J. Massague. Cancer metastasis: building a framework. Cell, 127 (2006), pp. 679–695.
[23]. D.A. Smirnov, D.R. Zweitzig, B.W. Foulk, M.C. Miller, C.V. Doyle, K.J. Piente, *et al.* Global gene expression profiling of circulating tumor cells. Cancer Res, 65 (2005), pp. 4993–4997.
[24]. R. Solmi, G. Ugolini, G. Rosati, S. Zanotti, M. Lauriola, I. Montroni, *et al*. Microarray-based identification and RT-PCR test screening for epithelial-specific mRNAs in peripheral blood of patients with colon cancer. BMC Cancer, 6 (2006), p. 250.
[25]. R. Solmi, P. De Sanctis, C. Zucchini, G. Ugolini, G. Rosati, M. del Governatore, *et al*. Search for epithelial-specific mRNAs in peripheral blood of patients with colon cancer by RT-PCR. Int J Oncol, 25 (2004), pp. 1049–1056.
[26]. S. Bonne, J. van Hengel, F. Nollet, P. Kools, F. van Roy. Plakophilin-3, a novel armadillo-like protein present in nuclei and desmosomes of epithelial cells. J Cell Sci, 112 (1999), pp. 2265–2276.
[27]. A. Schmidt, L. Langbein, S. Prätzel, M. Rode, H.R. Rackwitz, W.W. Franke. Plakophilin 3—a novel cell-type-specific desmosomal plaque protein. Differentiation, 64 (1999), pp. 291–306.
[28]. S. Bonné, B. Gilbert, M. Hatzfeld, X. Chen, K.J. Green, F. Van Roy. Defining desmosomal plakophilin-3 interactions. J Cell Biol, 161 (2003), pp. 403–416.
[29]. F.H. Brembeck, M. Rosario, W. Birchmeier. Balancing cell adhesion and Wnt signaling, the key role of beta-catenin. Curr Opin Genet Dev, 16 (2006), pp. 51–59.
[30]. S. Papagerakis, A.H. Shabana, J. Depondt, P. Gehanno, N. Forest. Immunohistochemical localization of plakophilins (PKP1, PKP2, PKP3, and p0071) in primary oropharyngeal tumors: correlation with clinical parameters. Hum Pathol, 34 (2003), pp. 565–572.
[31]. C. Furukawa, Y. Daigo, A. Takano, N. Ishikawa, T. Kato, S. Hayama, *et al*. Plakophilin 3 oncogene as prognostic marker and therapeutic target for lung cancer. Cancer Res, 65 (2005), pp. 7102–7110.
[32]. J. Schwarz, A. Ayim, A. Schmidt, S. Jäger, S. Koch, R. Baumann, *et al*. Differential expression of desmosomal plakophilins in various types of carcinomas: correlation with cell type and differentiation. Hum Pathol, 37 (2006), pp. 613–622.
[33]. Valladares-Ayerbes M, Díaz S, Medina V, Iglesias P, Lorenzo MJ, Santamarina I, et al. Evaluation of plakophilin 3 as a molecular marker for micrometastasis in gastrointestinal cancer. In: Grunbag SM, editor. Proceedings of the gastrointestinal cancer symposium; January 19–21, 2007; Orlando, USA. American Society of Clinical Oncology; 2007. p. 302 [Abstract 438].
[34]. K. Aigner, L. Descovich, M. Mikula, A. Sultan, M. Schreiber, W. Mikulits, *et al.* The transcription factor ZEB1 (dEF1) represses plakophilin 3 during human cancer progression. FEBS Letters, 581 (2007), pp. 1617–1624.
[35]. U. Woelfle, J. Cloos, G. Sauter, L. Riethdorf, F. Janicke, P. Van Diest, *et al*. Molecular signature associated with bone marrow micrometastasis in human breast cancer. Cancer Res, 63 (2003), pp. 5679–5684.
[36]. J.-S. Zhang, A. Gong, J.C. Cheville, D.I. Smith, C.Y.F. Young. AGR2, an androgen-inducible secretory protein overexpressed in prostate cancer. Genes Chromosomes Cancer, 43 (2005), pp. 249–259.

[37]. F.R. Fritzsche, E. Dahl, S. Pahl, M. Burkhardt, J. Luo, E. Mayordomo, *et al*.. Prognostic relevance of AGR2 expression in breast cancer. Clin Cancer Res, 12 (2006), pp. 1728–1734.
[38]. S. Lee, S. Bang, K. Song, I. Lee. Differential expression in normal-adenoma-carcinoma sequence suggests complex molecular carcinogenesis in colon. Oncol Rep, 16 (2006), pp. 747–754.
[39]. H. Hong Zhu, D.C. Lam, K.C. Han, V.P. Tin, W.S. Suen, E. Wang, *et al*. High resolution analysis of genomic aberrations by metaphase and array comparative genomic hybridization identifies candidate tumour genes in lung cancer cell lines. Cancer Lett, 245 (2007), pp. 303–314.
[40]. E. Pohler, A.L. Craig, J. Cotton, L. Lawrie, J.F. Dillon, P. Ross, *et al*. The Barrett's antigen anterior gradient-2 silences the p53 transcriptional response to DNA damage. Mol Cell Proteomics, 3 (2004), pp. 534–547.
[41]. D. Liu, P.S. Rudland, D.R. Sibson, A. Platt-Higgins, R. Barraclough. Human homologue of cement gland protein, a novel metastasis inducer associated with breast carcinomas. Cancer Res, 65 (2005), pp. 3796–3805.