Am J Transl Res 2020;12(9):5797-5807
www.ajtr.org /ISSN:1943-8141/AJTR0100791

Original Article

Prognostic impact of cancer stem cell markers *ABCB1*, *NEO1* and *HIST1H2AE* in colorectal cancer

Bogdan Badic[1], Stéphanie Durand[2*], Flaria El Khoury[2&], Pierre De La Grange[4], David Gentien[3], Brigitte Simon[2], Catherine Le Jossic-Corcos[2#], Laurent Corcos[2#]

[1]INSERM UMR 1101, [2]INSERM UMR 1078, Université de Brest, 22 Avenue Camille Desmoulins, 29238 Brest, France; [3]Institut Curie, Département de Recherche Translationnelle, Plateforme Génomique, 1 Avenue Claude Vellefaux, 75010 Paris, France; [4]GenoSplice Technology, 19 Rue Claude Bernard, 75005 Paris, France; [&]Present address: CNRS UMR 9197, Neuro-PSI, 1 Avenue de la Terrasse, 91198 Gif-Sur-Yvette, France; [*]Present address: EA7500, Université de Limoges, Faculté des Sciences et Techniques, 123 Avenue Albert Thomas, 87060 Limoges, France. [#]Equal contributors.

Received August 10, 2019; Accepted March 23, 2020; Epub September 15, 2020; Published September 30, 2020

Abstract: Colon cancer develops according to a defined temporal sequence of genetic and epigenetic molecular events that may primarily affect cancer stem cells. In an attempt to identify new markers of such cells that would help predict patient outcome, we performed a comparative transcriptome analysis of colon cancer stem cells and normal colon stem cells. We identified 162 mRNAs, either over- or under-expressed. According to Cox multivariate regression with our set of 83 colorectal cancers, low expression of *ABCB1, NEO1,* tumor size and the presence of distant metastases were predictive factors for overall survival. Combined expression of *ABCC1* and *NEO1* was a significant predictor for overall survival in our cohort, which was confirmed by external validation in 221 colorectal cancers from the Cancer Genome Atlas (TCGA) portal. Tumor size, lymph node involvement and *HIST1H2AE* expression were also independently correlated with disease-free survival. Taken together, our results suggest that molecular markers of colorectal cancers *ABCB1, NEO1* and *HIST1H2AE* are prognostic factors in colorectal cancer patients. It can be proposed that surveying expression of these marker genes should help better characterizing CRC prognosis, and help selecting the best therapeutic options.

Keywords: Colorectal cancer, cancer stem cell, *ABCB1*, *NEO1*, *HIST1H2AE*, prognostic factors, whole transcriptome analysis

Introduction

Colorectal cancer (CRC) is the third most common cancer in men and the second in women worldwide [1]. Incidence is low before the age of 50, but strongly increases thereafter. Median age at diagnosis is about 70 in developed countries [2]. CRC has a good prognosis when diagnosed at an early stage: the 5-year relative survival is 91% for localized stages and 70% in case of loco-regional invasion [3]. However, 5-year survival drops down to roughly 11% in metastatic situations, which represent approximately 25% of patients at diagnosis [4]. The death rate from CRC has decreased over the last 20 years, thanks to better disease management (early diagnosis and improvement of therapeutic modalities), but remains devastating worldwide [5].

Chemotherapy is the standard treatment for metastatic colorectal cancer as it prolongs survival and improves life quality [6]. Whereas objective response rates were only 20 to 30% with the combination of 5-fluoro-uracile/folinic acid, the addition of new drugs, such as oxaliplatin or irinotecan, helped increasing response rates up to around 50% and improved median survival from 6 months to about 2 years [7, 8]. The emergence of targeted therapies, such as Epidermal Growth Factor (EGFR) receptor inhibitors (Cetuximab) or angiogenesis inhibitors (Bevacizumab) improved these figures further [9]. However, the 5-year survival rate of patients treated with chemotherapy alone remains below 1% [10]. In this context, it is imperative to better comprehend the developmental process of cancer lesions and to move towards more individualized treatments.

It is believed that colon stem cells from the bottom of the intestinal crypts can undergo molecular changes that make them turn into a "cancer-founder" cell population resistant to conventional anticancer therapies [11]. Nevertheless, the distinction between normal and cancer stem cells is not well defined, and it is difficult to recognize cell markers specific to the cancer-initiating population. According to the cancer stem cell hypothesis for the origin of aggressive CRCs, surveying such markers could be a tractable way to predict cancer outcome [12]. Although this could be achieved on a per case basis, by analyzing the putative links of candidate genes with colorectal cancer, a blind, transcriptome-wide survey would address this question in a more comprehensive way. In the present study, we conducted a differential transcriptome analysis between stem cell populations derived from normal colons and from cancerous colons. This approach led us to sort out several markers specific for the cancer stem cell compartment. We analyzed expression of these markers in colorectal tumors, together with that of already reported CRC markers, and confronted these data to histopathology characteristics and patient survival. Using this approach, we identified novel molecular markers associated with either disease-free (DFS) or overall survival (OS).

Materials and methods

Gene-expression profiling

Stem cell populations from either disease-free colon tissues or colon cancers were purchased from Celprogen™ (3 populations of each, Torrance, CA) and grown according to the manufacturer's instructions. Stem cells were from male donors of Caucasian origin with a mean age of 60. Total RNA was extracted with TRIzol® reagent (ThermoFisher scientific, Illkirch, France) according to the manufacturer's instructions and quality controls were performed using the RNA 6000 Nano LabChip® and the 2100 BioAnalyzer (Agilent Technologies, Massy, France). GeneChip® Human Transcriptome Array 2.0 (HTA2.0, Affymetrix, Santa Clara, CA, USA) hybridization was performed at the Curie Institute microarray core facility (Paris, France) according to Affymetrix procedures. Affymetrix HTA2.0 dataset analysis and visualization were performed using EASANA® (GenoSplice technology), based on the GenoSplice's FAST DB® 2014.1 annotations [13]. Transcriptomic data were normalized using quantile normalization. Background corrections were made with anti-genomic probes selected as described [14, 15]. Only probes targeting exons annotated from FAST DB® transcripts were selected to focus on well-annotated genes whose mRNA sequences were in public databases. Bad-quality selected probes (*e.g.* probes named by Affymetrix as 'cross-hybridizing') and probes whose intensity signal was too low compared to anti-genomic background probes with the same GC content were removed from the analysis. Only probes with a Detection Above BackGround (DABG) p-value ≤ 0.05 in at least half of the arrays were considered for statistical analysis. Only genes expressed in at least one compared condition were analyzed. Gene expression was recorded only if at least half of the gene probes had a DABG p-value ≤ 0.05. Unpaired Student's t-tests were performed to compare gene intensities in the different biological replicates. Genes were considered as significantly regulated when fold-changes were ≥ 1.5 and raw p-values ≤ 0.05.

Real-time PCR analyses

To evaluate the prognostic impact of the transcript content in CRC compared to healthy tissue, the differential expression of 21 selected genes was analyzed by real-time PCR (StepOne plus, Applied Biosystems). Briefly, total RNA was extracted using NucleoSpin® RNA kit (Macherey-Nagel, Hoerdt, France) according to the manufacturer's instructions. Complementary DNA synthesis and real-time PCR were performed as described [16]. All conditions were normalized relative to the RPLP0 (ribosomal protein P0) control RNA. PCR primers were purchased from Eurogentec (Seraing, Belgium). Primer sequences will be made available on request. The results were analyzed using the ΔΔCt method [17].

Patients and clinical data

All patients signed an informed consent form in which they granted use of data obtained from analysis of their tissue samples. Eighty-three patients from our institution were included into the study: 36 women (43%) and 47 men (57%), with an average follow-up of 32.7 (1-117) months. The average age of the patients was 71 (26-94) with an average American

Society of Anesthesiologists physical status score of 2.64. Tumors were localized in the right colon in 38 (46%) patients, in the left colon in 29 (35%) patients, and in the rectum in 16 (19%) patients. The preoperative extension assessment was positive with the discovery of synchronous liver lesions for 13 (16%) patients and pulmonary lesions for 7 (8%) patients.

External validation cohort

We evaluated the prognostic value of our results in a public database combining tumor gene expression and outcome information for 221 patients from The Cancer Genome Atlas (TCGA) portal. Clinical information and microarray normalized expression data of the COADREAD cohort (153 colon and 68 rectal carcinomas as compared to 22 NT) [18] were downloaded from the Firehose portal of the Broad Institute. The distribution of grading stages of TCGA colorectal carcinoma was: I: 47, II: 86, III: 54 and IV: 34.

Statistical analysis

Statistical analyses were performed using MedCalc Statistical Software version 14.8.1 (MedCalc Software, Ostend, Belgium). Comparisons between gene expression profiles from microarray analyses and continuous variables were performed using the non-parametric Kruskal-Wallis test. Comparisons between gene expression profiles from real-time PCR and discrete variables were performed using the Chi2 test or Fisher's exact test. DFS was defined as the time from diagnosis to first event (local or metastatic failure or death). DFS was investigated only for 66 patients (stages I to III); already metastatic patients at time of diagnosis were excluded. Patients with no events were censored at time of last follow-up. OS was defined as the time from diagnosis to death from any cause or last follow-up, for all 83 patients. The prognostic value of each feature for outcome was assessed using the Kaplan-Meier method and log-rank test with cut-off thresholds determined by receiver operating characteristics curve (ROC) analysis, according to Youden's index [19]. For each variable, relative risks were estimated using a univariate Cox model and expressed with their 95% confidence interval. Multivariate analysis was carried out using a Cox regression model. A *p*-value below 0.05 was considered as significant. All clinical and gene expression data were used in a multivariate Cox regression analysis. In order to evaluate the improvement in prognosis stratification, models combining genes expression with their optimal cut-off values were tested among the two cohorts. The resulting Kaplan-Meier curves were compared using median OS in each group, hazard ratios (HR) and associated 95% confidence intervals (CI). Higher values of HR, with 1 being excluded of the 95% CI, indicated models with better stratification power.

Results

Transcriptome-wide gene expression analysis identifies differences between colon cancer stem cells and colon stem cells

Based on the hypothesis that colon cancer stem cells should show variations in gene expression as compared to healthy colon stem cells, we used three distinct stem cell populations derived from colon cancers and three distinct stem cell populations isolated from normal colons. We performed differential gene expression analysis with Affymetrix GeneChip® Human Transcriptome Array 2.0. Principal component analysis and heatmap clustering based on sample-to-sample distances showed a clear distinction of the two populations of cells, although the normal stem cell populations were somewhat heterogenous (Supplementary Figure 1A and 1B). One hundred and sixty-two genes were differentially expressed (> 1.5-fold up- or down-regulation), including 77 up-regulated genes (48%) and 85 down-regulated genes (52%) (Supplementary Table 1). Hierarchical clustering showed a clear separation of the two sets of cells (Supplementary Figure 1C). In addition, functional gene enrichment using Gene Ontology and KEGG pathway analyses revealed enrichment in gene sets involved in transcriptional regulation by nucleosome, telomere and chromatin assembly and, to a lesser extent, in FoxO and SMAD transduction signals, response to drugs or glutathione metabolism (Supplementary Table 2).

Real-time PCR analysis of selected genes uncovers differences between colon cancer and colon cancer stem cells

We performed real-time PCR experiments with our cohort of CRC (n=83) samples to analyze

expression of a set of colon cancer stem cell markers identified through our microarray analysis, together with colon cancer stem cell-like markers that we have identified previously (*ABCB1, ABCC2, ABCG2, ALDH1A1, CD166* and *CDKN1A*) [20]. Whereas expression of *AQP1, PKIA, HIST1H3J* and *LINC00883* was increased and that of *MAPK10* was decreased in CRC and cancer stem cell samples, as compared to normal colon or stem cells, respectively, three genes (*INHBB, UQCC2* and *NEO1*) showed opposite variation. Expression of the *HIST1H3F, HIST1H2AC, HIST1H2AE, SULF2* and *SMAD3* genes showed significant variation only in cancer stem cells *vs.* normal stem cells (**Table 1**). These results were compared to our previous gene expression profiling study in CRC [21] and to gene expression data obtained from microarray platforms and from the TCGA COADREAD cohort [18] (http://gdac.broadinstitute.org). The fold-changes obtained in these two external sets of CRC tumors were similar to those reported here (**Table 1**), confirming the soundness of our present gene expression analysis. The most up-regulated gene was *ABCC2* (6.75-fold), while the most suppressed gene was *ABCG2* (10-fold) (**Table 1**).

To determine if cancer stem cell markers could discriminate CRC samples, we performed a classification by hierarchical clustering analysis from the 21-gene expression data. Two CRC populations, noted "group A" (left, 47 CRC samples) and "group B" (right, 36 CRC) could be separated, according to differential gene expression patterns (**Figure 1**). However, we observed no clear association of this feature with histopathological characteristics of CRC (TNM or stage) although group A cancers contained the main part of dead patients. In fact, Kaplan-Meier survival analysis showed that group A patients had the worst prognosis with a mean survival time of 55.8 ± 6.8 months (median survival: 60 months), whereas group B patients had a better prognosis with a mean survival time of 82.1 ± 8.5 months (median survival not available) (P=0.027, log rank test) (**Figure 2A**). The heatmap highlighted global gene expression levels specific to each group and, for a number of them, the mean expression levels were strongly affected. Indeed, *ABCC2, ABCB1, ABCG2, AQP1, INHBB, LINC00883* or *MAPK10* showed a ratio of means of groups B and A above 2.5-fold (**Figure 2B**). Other genes, such as histone family genes, showed only mild differences between the two groups with a ratio ranging from 1.3- to 1.6-fold.

Correlation between patient outcome and gene expression

CRC classification, based on our 21-gene set, indicated that we could identify potential new independent CRC prognostic markers. In order to look for associations of such markers and patient survival or death, we performed Kaplan-Meier analysis by univariate and multivariate Cox regression for each marker and histopathological characteristics for our 83 CRC cohort and the external 221-CRC cohort from TCGA. Univariate analysis showed significant correlations for *ALDH1*, *LINC00883*, and *ABCB1* genes expression with overall survival (OS). All clinical and gene expression data were used in a multivariate Cox regression analysis. *ABCB1, NEO1* gene expression, tumor size and the presence of distant metastases were prognostic factors for OS (**Table 2**). The Cox regression analysis using low expression of these 2 genes (*NEO1* < 0.50 and *ABCB1* < 0.50) showed a significant correlation with OS (OR 3.39 95% CI 1.41 to 8.16, P=0.0067). *NEO1* and *ABCC1* had a correlation rs < 0.7 and the combination of these two parameters led to a significant patient stratification (HR 3.74 95% CI 1.20 to 6.36, P=0.0072).

When considering the *ABCC1* and *NEO1* genes, together with their optimal cut-off values, for analyzing the TCGA cohort, we observed a significant correlation with OS (OR=1.95 95% CI=1.21 to 3.16, P=0.0062), as in our 83-cancer cohort. Combined expression of these genes was an independent risk factor for OS (HR 1.95 95% CI 1.19 to 3.22, P=0.005) (**Figure 3**). In addition, DFS survival was correlated with tumor (P=0.0001) and node status (P=0.0124) histopathological factors on univariate regression. Cox multivariate analysis found *HIST1H2AE* expression and tumor and node status to be correlated with DFS. DFS in the external TCGA cohort was not analyzed because the corresponding data were not available.

Discussion

The stem cell hypothesis of cancer, and especially CRC, has been amply substantiated [22,

Table 1. Gene expression in colon cancer samples versus normal samples

Comparison	CRC *vs.* NT	CRC *vs.* NT		CRC *vs.* NT		CRA *vs.* NT		Colon Cancer Stem Cells *vs.* Normal Stem Cells	
RNA quantification method	Real-time PCR	Custom Agilent 244K Gene Expression Microarray		Agilent 44k Whole Human Genome microarrays		Agilent 44k Whole Human Genome microarrays		Affymetrix HTA 2.0 microarray	
Source	Present study	The Cancer Genome Atlas, Nature 2012		GSE50117, Pesson *et al*., PLoS One 2014		GSE50114, Pesson *et al*., PLoS One 2014		Present study	
Gene Symbol	Fold-Change	Fold-Change	*p*-adj (BH)	Fold-Change	*p*-adj (BH)	Fold-Change	*p*-adj (BH)	Fold-Change	*p*-value
ABCC2	**6.75**	**3.45**	0.0007	/	/	/	/	/	/
INHBB	**3.14**	**4.22**	0.00008	**2.53**	0.025	/	/	**0.625**	0.012
AQP1	**2.37**	**0.57**	3.6×10^{-8}	**0.92**	n.s	**0.24**	0.0001	**2.16**	0.002
PKIA	**2.14**	0.89	0.008	**0.60**	n.s	**0.6**	n.s	**2.67**	0.0007
CD166	**1.85**	**1.89**	3.6×10^{-6}	**2.28**	0.003	/	/	/	/
HIST1H3J	**1.85**	0.87	0.008	1.12	n.s	0.82	n.s	**1.93**	0.034
LINC00883	**1.76**	/	/	/	/	/	/	**1.57**	0.042
UQCC2	**1.53**	**1.68**	1.8×10^{-6}	**1.86**	0.0007	**2.09**	6.1×10^{-6}	**0.52**	0.03
HIST1H3F	1.37	0.95	n.s	**1.6**	0.0036	1.27	n.s	**1.98**	0.036
HIST1H2AC	1.33	0.95	n.s	0.79	n.s	0.95	n.s	**1.97**	0.009
HIST1H2AE	1.26	0.87	0.01	1.21	n.s	0.97	n.s	**2.27**	0.041
SULF2	1.13	1.1	n.s	1.22	n.s	0.69	0.08	**0.38**	0.0005
DCN	0.95	**0.41**	5.5×10^{-11}	**0.5**	0.01	**0.12**	8.4×10^{-10}	**0.55**	0.002
SMAD3	0.81	**0.65**	4.6×10^{-10}	0.82	n.s	1.06	n.s	**2.18**	0.00002
CDKN1V1	**0.66**	**0.58***	3.5×10^{-8}	**0.6***	0.015	1.42*	n.s	/	/
ALDH1A1	**0.65**	**0.58**	4.9×10^{-6}	**0.41**	0.07	0.82	n.s	/	/
CDKN1V2	**0.59**	**0.58***	3.5×10^{-8}	**0.6***	0.015	1.42*	n.s	/	/
ABCB1	**0.53**	**0.33**	2.2×10^{-11}	**0.26**	0.047	**0.48**	0.05	/	/
NEO1	**0.49**	**0.48**	3.2×10^{-14}	**0.38**	0.00003	**0.65**	0.037	**2.3**	0.001
MAPK10	**0.48**	0.74	0.0002	/	/	/	/	**0.29**	0.00002
ABCG2	**0.10**	**0.07**	2.7×10^{-44}	**0.05**	1.3×10^{-7}	**0.05**	1.6×10^{-14}	/	/

Real-time PCR experiments were performed with all individual colon carcinoma RNA samples from our set (n=83). To facilitate multi-platforms comparisons, we have indicated i) fold-change of CRC *vs.* NT extracted from TCGA microarray (221 CRC *vs.* 22 NT) data (https://gdac.broadinstitude.org); ii) fold-change from our previous microarray study of CRC (n=9) and CRA (n=37) as compared to normal mucosae (n=9) (Pesson M. *et al*., PLoS One 2014) (GSE50117 and GSE50114 GEO dataset accession number, https://www.ncbi.nlm.nih.gov/gds); iii) fold-change of cancerous *vs.* normal colon stem cells. The *p*-value corresponds to Student t-test comparison with Benjamini-Hochberg (BH) correction for multiple hypothesis testing. Bold names and values were above set thresholds (> 1.5 fold up- or down-regulation), applied for differential gene expression identification in colon cancer stem cells versus normal stem cells. *Fold-change for all transcripts without discrimination. CRC, colorectal adenocarcinoma; CRA, colorectal adenoma; NT, normal colon tissue. /, missing information; n.s, not significant.

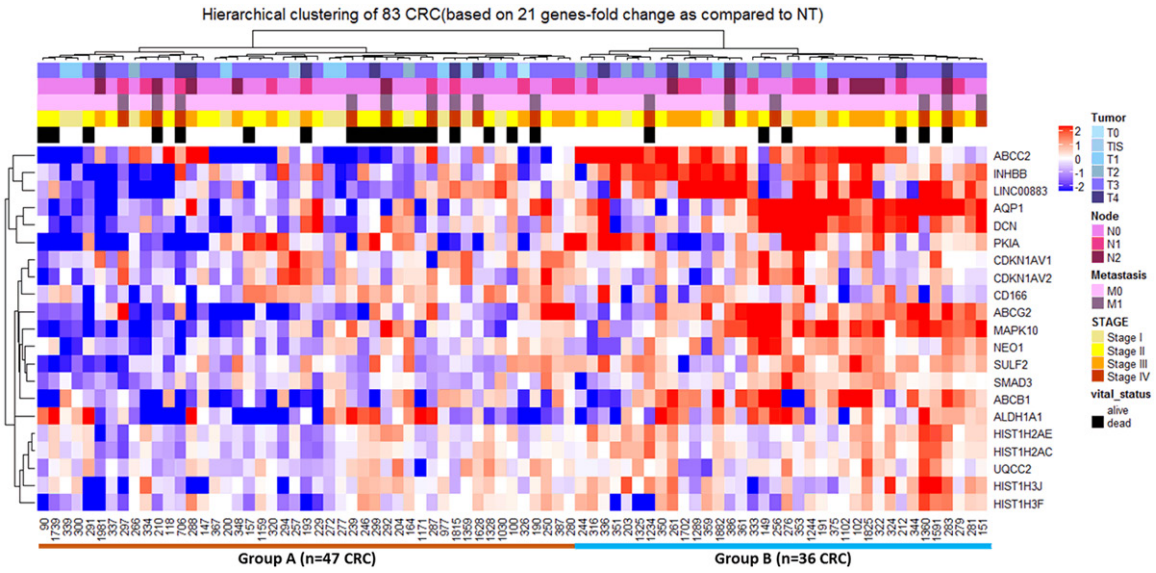


Figure 1. Hierarchical clustering of 83 CRC with the 21-gene set. Log_2-transformed and median-centered fold-change values obtained by comparison of CRC and normal colon tissue expression levels were hierarchically clustered using Euclidean distance measure and Ward method linkage. Red and blue colors indicate transcript levels above and below the mean values, respectively. Tumor samples are identified by a number and genes are identified by their symbols. Clinicopathological information (pTNM, stage and vital status) for CRC is indicated at the top of the heatmap. Each column shows the gene expression profile of a sample, and each line shows the variations in the expression level of a given gene among tumor samples. The length of the branches on the trees forming the dendrogram on the top of each panel reflects the degree of similarity between samples. Subdivision of samples into two groups (A and B) according to the dendrogram is used for survival analysis.

23]. According to this hypothesis, intestinal stem cells may undergo genetic and epigenetic changes that are transferable, at least partly, to the progeny of differentiated cells [24]. As a result, cancer cells acquire a selective growth advantage over normal epithelial cells, proliferate and eventually metastasize to distant organs. One way to distinguish the most aggressive cancers, *i.e.* prone to colonize distant tissue sites, from less aggressive cancers, is to use specific expression markers of this sub-population of cancer-initiating cells. To identify such markers, we compared colon stem cells populations originating from either cancer or healthy colon tissues. Our DNA chip experiments identified 162 markers that distinguished these cell populations (threshold mean expression value difference of 1.5-fold). Several marker genes were distinctive, showing either increased (*PKIA*, *NEO1*, *SMAD3*, *HIST1H3B, HIST1H3J* or *LINC00883*) or decreased (*MAPK10, SULF2* or *UQCC2*) expression between stem cell populations. Adding several already reported colon cancer markers to the stem cell markers, we engineered a PCR assay to survey expression of all the genes at once in colon cancer *vs.* cancer-free colon tissue, and confronted their expression variation to patient outcome.

ABC transporter genes were the most affected, *ABCC2* being the most overexpressed gene and *ABCG2* the most down-regulated gene, in agreement with previous reports [25]. Quite strikingly, every single colon cancer sample from our cohort showed significantly reduced expression of *ABCG2,* suggesting that the functions endorsed by this gene are strongly detrimental for cancer development. Importantly, ABC proteins participate in exchanges between the intra- and extracellular compartments, indicating that one strong characteristic of colon cancer is a modification of interactions of tumor cells with neighboring cells and/or with extracellular components. In addition, *ABCC2* over-expression has been linked to cisplatin resistance [26] and *ABCG2* overexpression was associated with multiple drug resistance [27, 28], but its reduction is also likely to restrict the protection capacity of the cells against environmental xenobiotics [29]. However, in our study, colon cancer stem cells

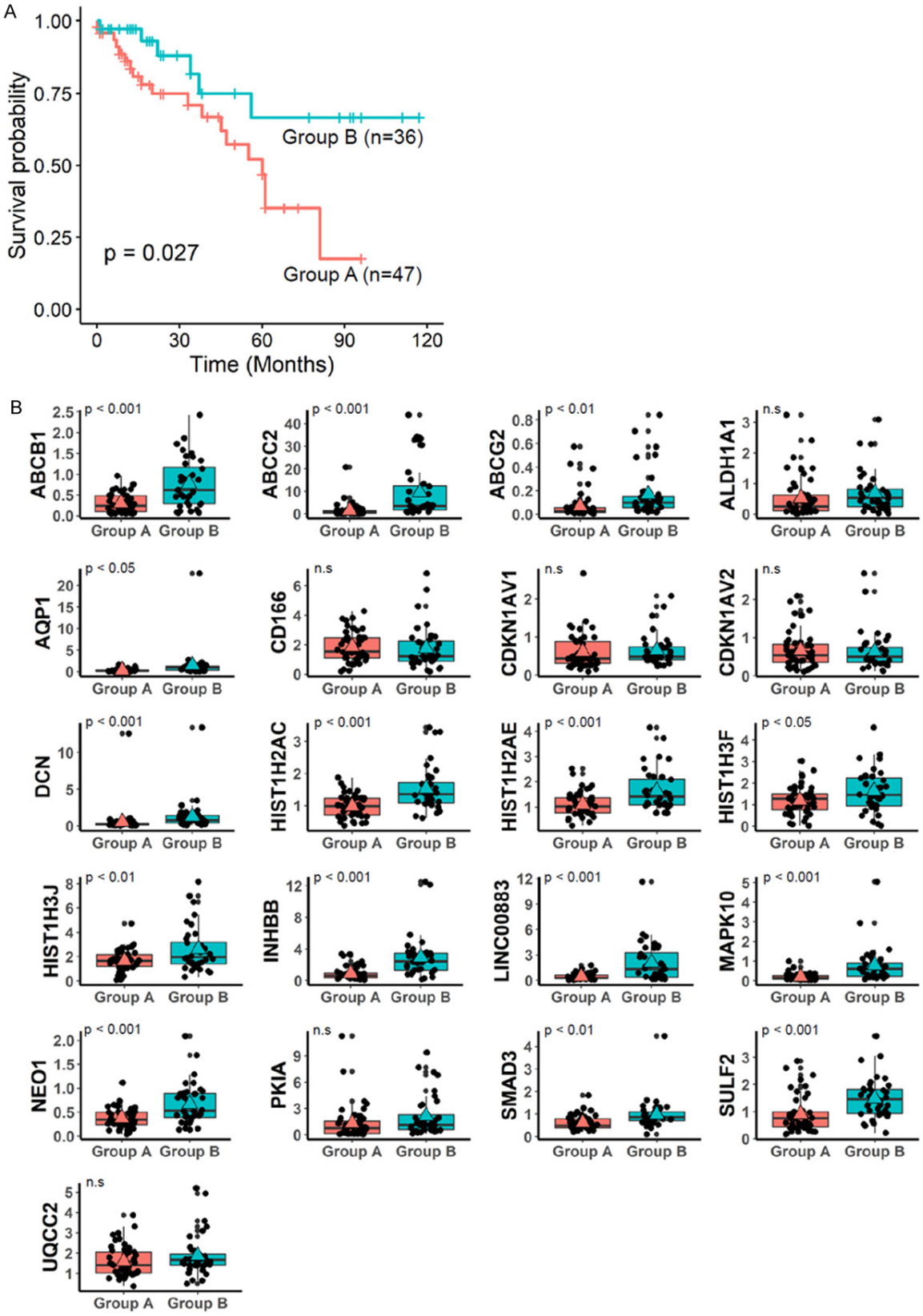


Figure 2. A. Overall survival in the 83 CRC cohort. Kaplan-Meier analysis compared survival of two groups defined by hierarchical clustering according to the expression level of our 21-genes set (**Table 1**). Group A (orange curve, n=47

CRC) defined the poor survival group and group B (blue curve, n=36 CRC) defined a better survival group (p=0.027, log-rank test). B. Gene-expression level comparison between poor and better overall survival groups of CRC. Box plots displaying the fold-change obtained by comparison of CRC and NT expression values evaluated by quantitative RT-PCR for the A (orange box plot) and B (blue box plot) groups. The average fold-change for each group is indicated by a triangle symbol. *P*-values were obtained by comparison of A and B groups by unpaired t-test. There was no significant difference between poor and better survival groups with respect to the expression level of *ALDH1A1*, *CD166*, *CDKN1AV1*, *CDKN1AV2*, *PKIA* and *UQCC2*.

Table 2. Overall survival and disease-free survival Cox multivariate analysis

	Covariate	OR	95% CI of OR	*p*-value
Overall survival				
Histopathological	T4	4.4575	1.3813 to 14.3850	0.0129
Gene expression	M1	7.6877	2.8501 to 20.7364	0.0001
	NEO1	28.0081	3.5036 to 223.9015	0.0018
	ABCB1	0.0168	0.0021 to 0.1356	0.0001
	NEO1 and *ABCB1**	3.3914	1.4097 to 8.1592	0.0067
Disease-free survival				
Histopathological	T4	668.4886	7.449 to 59987.954	0.0048
Gene expression	N1	82.4771	3.142 to 2164.529	0.0085
	HIST1H2AE	4.8734	1.043 to 22.765	0.0451

T: tumor size; M: metastasis; N: lymph node; * model combining low expression of *NEO1* (< 0.50) and *ABCB1* (< 0.50). All clinical and gene expression data were used in a multivariate Cox regression analysis. Prognostic value of each feature for outcome was assessed using the Kaplan-Meier method and log-rank test with cut-off thresholds determined by receiver operating characteristics curve (ROC) analysis, according to Youden's index [19]. For each variable, relative risks were estimated using a univariate Cox model and expressed with their 95% confidence interval. Multivariate analysis was carried out using a Cox regression model. A *p*-value below 0.05 was considered as statistically significant.

were not separated from colon stem cells based on expression of these transporter genes (above the 1.5-fold threshold), or on expression of other genes that nevertheless, were either over- or under-expressed, or even unchanged in CRC. *CD166*, *HIST1H3J* and *PKIA* were distinctive both of CRC *vs.* normal tissue and colon cancer stem cells *vs.* colon stem cells. Therefore, at the level of our population of cancer patients, we may suggest that these genes could be fair indicators of cancer aggressiveness. By contrast, the variation in expression of *INHBB* and *MAPK10* was opposite between CRCs and colon cancer stem cells. In addition, expression of *SMAD3* was barely reduced in colon cancer, but increased in colon cancer stem cells, and expression of the *SULF2* gene was decreased in colon cancer stem cells, but unchanged in colon cancer. These similarities, as well as these differences, between colon cancers and colon cancer stem cells suggest that analysis of CRC as a whole cannot be fully recapitulated by the sole analysis of gene expression in colon cancer stem cells. However, these results point to the importance of understanding the individual roles played by those genes, in particular *LINC00883*, which is highly expressed in pluripotent stem cells, and represses the *DPPA2* muscle-specific gene [30].

We identified several correlations between CRC patient outcome and gene expression alterations in colon cancer stem cells. *HIST1H2AE* expression was an independent prognostic factor of DFS, whereas *ALDH1A1*, *LINC00883, ABCB1* and *NEO1* expression were prognostic factors of OS. Interestingly, *ALDH1A1* has been proposed to define a subpopulation of tumor initiating cells or as a marker of stem cells present in different cancers [31-33].

We found that combined expression of *ABCC1* and *NEO1* was an important predictor of overall survival in CRC patients. Lower *ABCB1* expression was already found in poorly differentiated CRC tumors in line with studies in cell lines, where *ABCB1* had a higher expression in well-differentiated colon cancer cells compared to poorly differentiated cells [34-36]. Neogenin 1 (*NEO1*) is a receptor of the Deleted in Colorectal Carcinoma (DCC)/Frazzled/UNC-40 family, which regulates axon guidance and stabilizes epithelial adherens junctions [37, 38]. The role

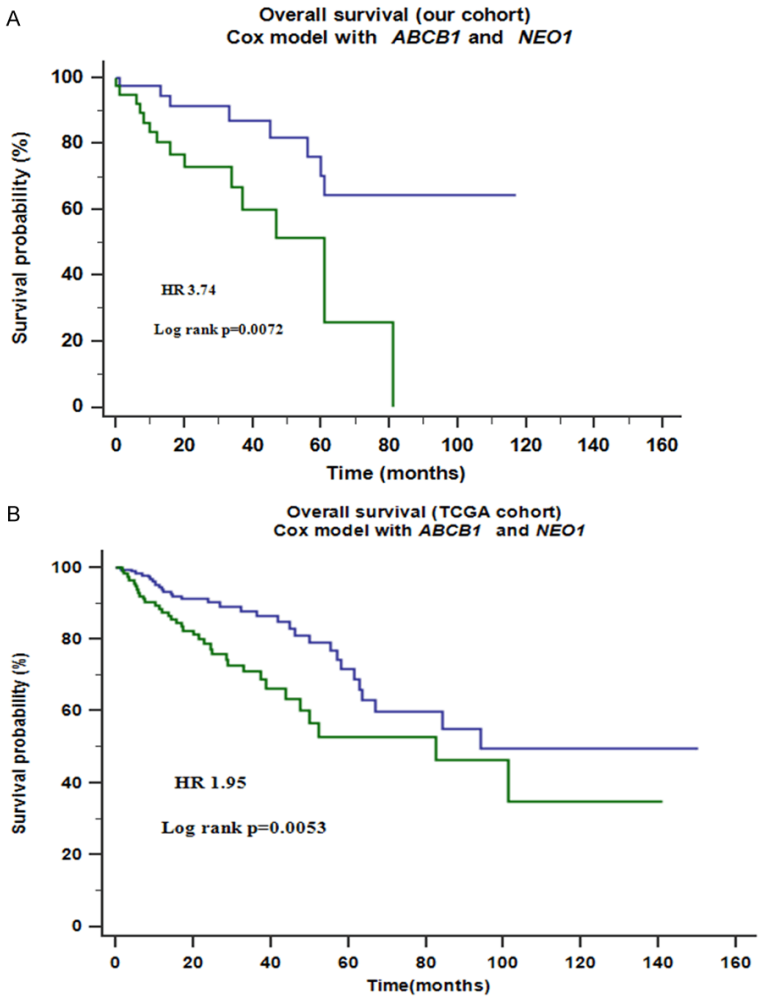


Figure 3. Kaplan Meier analysis of patient survival. All clinical and gene expression data were used in a multivariate Cox regression analysis. Prognostic value of each feature for outcome was assessed using the Kaplan-Meier method and log-rank test with cut-off thresholds determined by receiver operating characteristics curve (ROC) analysis, according to Youden's index [19]. The analysis was performed upon combining low expression of *NEO1* (fold-change < 0.50) and *ABCB1* (fold-change < 0.50). A. Our cohort (HR 3.74 log rank *P*=0.0072). B. The TCGA cohort (HR 1.95 log rank *P*=0.0053).

of neogenin 1 in maintaining adherens junctions and its loss in carcinomas may contribute to metastasis by promoting EMT and increasing motility [39].

Taken together, our results identified gene markers specific for colon cancer stem cells, some of which were also specific for colon cancers, while others were deregulated in an opposite way. We also observed correlations between altered *ABCB1* and *NEO1* gene expression levels and patient outcome. In the future, it could be advisable, in addition to these genes, to also survey the expression of colon cancer stem cell markers, such as *CD166*, *HIST1H3J* and *PKIA* at time of patient diagnosis. Indeed, since these markers showed a similar trend of variation in CRC and in colon cancer stem cells, as compared to healthy colon or normal colon stem cells, respectively, their systematic analysis should help to better adapt the treatment and monitoring of patients with colorectal tumors. How these genes participate in the steps that lead to cancers, or, possibly, that oppose cancer development will have to be further investigated.

One limitation of our study is that it is most probable that our marker selection was not exhaustive, but this limitation was intrinsic to the fact that we purposely compared stem cell populations from normal and cancerous colons in the first place. In future experiments, it would be important to analyze expression of these genes in prospectively recruited colon cancer patients to see to what extent these markers might be, individually or combined, predictive of patient survival or death.

Acknowledgements

We thank Dr. Laurent Doucet for providing colon cancer tissue and Dr. Marie-Cécile Nicot for providing tissue fragments from Brest tumor tissue bank. This study was supported in part by the Ligue contre le Cancer, comité du Finistère and the INSERM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosure of conflict of interest

None.

Address correspondence to: Bogdan Badic, INSERM UMR 1101, Université de Brest, 22 avenue Camille Desmoulins, 29238 Brest, France. E-mail: bogdan.badic@chu-brest.fr

References

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D and Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015; 136: E359-386.

[2] Brenner H, Kloor M and Pox CP. Colorectal cancer. Lancet 2014; 383: 1490-1502.

[3] Haggar FA and Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clin Colon Rectal Surg 2009; 22: 191-197.

[4] Siegel R, DeSantis C, Virgo K, Stein K, Mariotto A, Smith T, Cooper D, Gansler T, Lerro C, Fedewa S, Lin C, Leach C, Cannady RS, Cho H, Scoppa S, Hachey M, Kirch R, Jemal A and Ward E. Cancer treatment and survivorship statistics, 2012. CA Cancer J Clin 2012; 62: 220-241.

[5] Edwards BK, Ward E, Kohler BA, Eheman C, Zauber AG, Anderson RN, Jemal A, Schymura MJ, Lansdorp-Vogelaar I, Seeff LC, van Ballegooijen M, Goede SL and Ries LA. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer 2010; 116: 544-573.

[6] Kopetz S, Chang GJ, Overman MJ, Eng C, Sargent DJ, Larson DW, Grothey A, Vauthey JN, Nagorney DM and McWilliams RR. Improved survival in metastatic colorectal cancer is associated with adoption of hepatic resection and improved chemotherapy. J Clin Oncol 2009; 27: 3677-3683.

[7] Jonker D, Rumble RB and Maroun J; Gastrointestinal Cancer Disease Site Groupof Cancer Care Ontario's Program in Evidence-Based Care. Role of oxaliplatin combined with 5-fluorouracil and folinic acid in the first- and second-line treatment of advanced colorectal cancer. Curr Oncol 2006; 13: 173-184.

[8] Kuebler JP and de Gramont A. Recent experience with oxaliplatin or irinotecan combined with 5-fluorouracil and leucovorin in the treatment of colorectal cancer. Semin Oncol 2003; 30: 40-46.

[9] Sanchez-Gundin J, Fernandez-Carballido AM, Martinez-Valdivieso L, Barreda-Hernandez D and Torres-Suarez AI. New trends in the therapeutic approach to metastatic colorectal cancer. Int J Med Sci 2018; 15: 659-665.

[10] Dy GK, Hobday TJ, Nelson G, Windschitl HE, O'Connell MJ, Alberts SR, Goldberg RM, Nikcevich DA and Sargent DJ. Long-term survivors of metastatic colorectal cancer treated with systemic chemotherapy alone: a North Central Cancer Treatment Group review of 3811 patients, N0144. Clin Colorectal Cancer 2009; 8: 88-93.

[11] Abdul Khalek FJ, Gallicano GI and Mishra L. Colon cancer stem cells. Gastrointest Cancer Res 2010; Suppl 1: S16-23.

[12] Shipitsin M and Polyak K. The cancer stem cell hypothesis: in search of definitions, markers, and relevance. Lab Invest 2008; 88: 459-463.

[13] de la Grange P, Dutertre M, Martin N and Auboeuf D. FAST DB: a website resource for the study of the expression regulation of human gene products. Nucleic Acids Res 2005; 33: 4276-4284.

[14] Gandoura S, Weiss E, Rautou PE, Fasseu M, Gustot T, Lemoine F, Hurtado-Nedelec M, Hego C, Vadrot N, Elkrief L, Letteron P, Tellier Z, Pocidalo MA, Valla D, Lebrec D, Groyer A, Monteiro RC, de la Grange P and Moreau R. Gene- and exon-expression profiling reveals an extensive LPS-induced response in immune cells in patients with cirrhosis. J Hepatol 2013; 58: 936-948.

[15] Schaller S, Buttigieg D, Alory A, Jacquier A, Barad M, Merchant M, Gentien D, de la Grange P and Haase G. Novel combinatorial screening identifies neurotrophic factors for selective classes of motor neurons. Proc Natl Acad Sci U S A 2017; 114: E2486-E2493.

[16] Durand S, Trillet K, Uguen A, Saint-Pierre A, Le Jossic-Corcos C and Corcos L. A transcriptome-based protein network that identifies new therapeutic targets in colorectal cancer. BMC Genomics 2017; 18: 758.

[17] Schmittgen TD and Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 2008; 3: 1101-1108.

[18] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012; 487: 330-337.

[19] Youden WJ. Index for rating diagnostic tests. Cancer 1950; 3: 32-35.

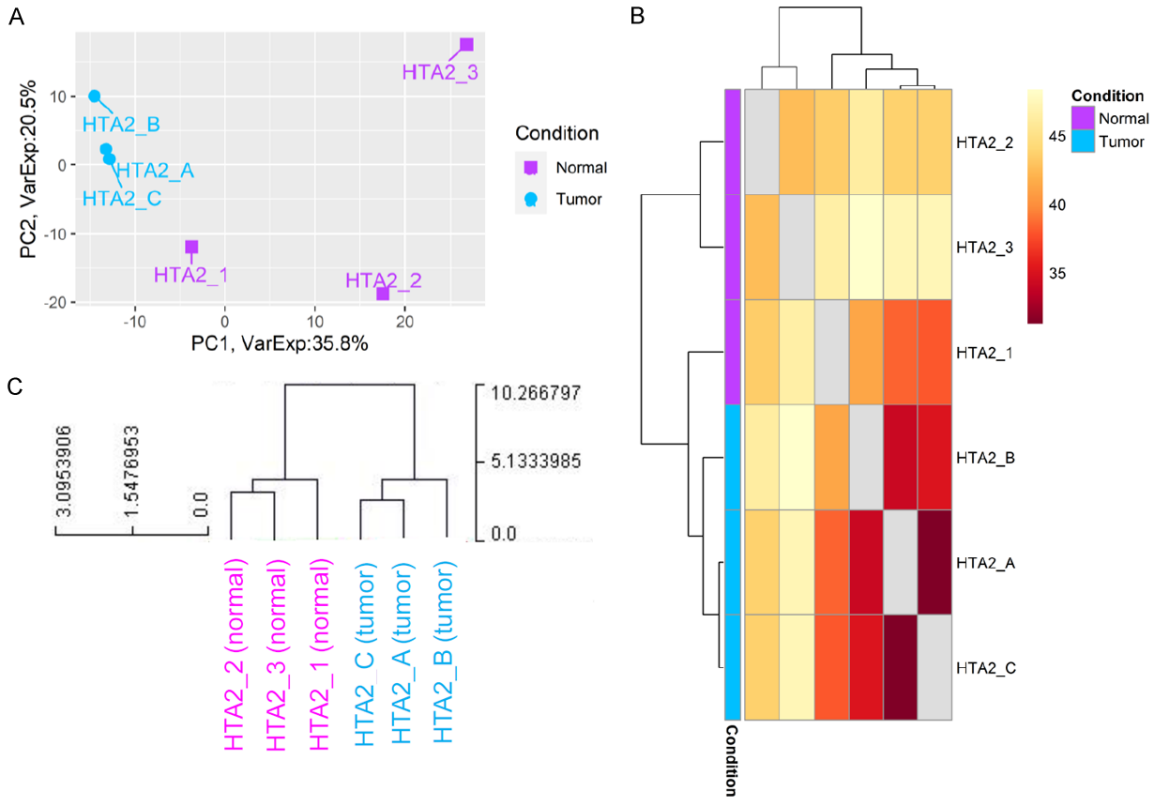
[20] El Khoury F, Corcos L, Durand S, Simon B and Le Jossic-Corcos C. Acquisition of anticancer drug resistance is partially associated with cancer stemness in human colon cancer cells. Int J Oncol 2016; 49: 2558-2568.

[21] Pesson M, Volant A, Uguen A, Trillet K, De La Grange P, Aubry M, Daoulas M, Robaszkiewicz M, Le Gac G, Morel A, Simon B and Corcos L. A gene expression and pre-mRNA splicing signature that marks the adenoma-adenocarcinoma progression in colorectal cancer. PLoS One 2014; 9: e87761.

[22] Tan BT, Park CY, Ailles LE and Weissman IL. The cancer stem cell hypothesis: a work in progress. Lab Invest 2006; 86: 1203-1207.

[23] Puglisi MA, Tesori V, Lattanzi W, Gasbarrini GB and Gasbarrini A. Colon cancer stem cells: con-

troversies and perspectives. World J Gastroenterol 2013; 19: 2997-3006.
[24] Ricci-Vitiani L, Lombardi DG, Pilozzi E, Biffoni M, Todaro M, Peschle C and De Maria R. Identification and expansion of human colon-cancer-initiating cells. Nature 2007; 445: 111-115.
[25] Andersen V, Vogel LK, Kopp TI, Saebo M, Nonboe AW, Hamfjord J, Kure EH and Vogel U. High ABCC2 and low ABCG2 gene expression are early events in the colorectal adenoma-carcinoma sequence. PLoS One 2015; 10: e0119255.
[26] Mazurowski MA. Radiogenomics: what it is and why it is important. J Am Coll Radiol 2015; 12: 862-866.
[27] Candeil L, Gourdier I, Peyron D, Vezzio N, Copois V, Bibeau F, Orsetti B, Scheffer GL, Ychou M, Khan QA, Pommier Y, Pau B, Martineau P and Del Rio M. ABCG2 overexpression in colon cancer cells resistant to SN38 and in irinotecan-treated metastases. Int J Cancer 2004; 109: 848-854.
[28] An Y and Ongkeko WM. ABCG2: the key to chemoresistance in cancer stem cells? Expert Opin Drug Metab Toxicol 2009; 5: 1529-1542.
[29] Nies AT and Keppler D. The apical conjugate efflux pump ABCC2 (MRP2). Pflugers Arch 2007; 453: 643-659.
[30] Wang L, Zhao Y, Bao X, Zhu X, Kwok YK, Sun K, Chen X, Huang Y, Jauch R, Esteban MA, Sun H and Wang H. LncRNA Dum interacts with Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. Cell Res 2015; 25: 335-350.
[31] Landen CN Jr, Goodman B, Katre AA, Steg AD, Nick AM, Stone RL, Miller LD, Mejia PV, Jennings NB, Gershenson DM, Bast RC Jr, Coleman RL, Lopez-Berestein G and Sood AK. Targeting aldehyde dehydrogenase cancer stem cells in ovarian cancer. Mol Cancer Ther 2010; 9: 3186-3199.
[32] Tomita H, Tanaka K, Tanaka T and Hara A. Aldehyde dehydrogenase 1A1 in stem cells and cancer. Oncotarget 2016; 7: 11018-11032.
[33] Yang L, Ren Y, Yu X, Qian F, Bian BS, Xiao HL, Wang WG, Xu SL, Yang J, Cui W, Liu Q, Wang Z, Guo W, Xiong G, Yang K, Qian C, Zhang X, Zhang P, Cui YH and Bian XW. ALDH1A1 defines invasive cancer stem-like cells and predicts poor prognosis in patients with esophageal squamous cell carcinoma. Mod Pathol 2014; 27: 775-783.
[34] De Iudicibus S, De Pellegrin A, Stocco G, Bartoli F, Bussani R and Decorti G. ABCB1 gene polymorphisms and expression of P-glycoprotein and long-term prognosis in colorectal cancer. Anticancer Res 2008; 28: 3921-3928.
[35] Hlavata I, Mohelnikova-Duchonova B, Vaclavikova R, Liska V, Pitule P, Novak P, Bruha J, Vycital O, Holubec L, Treska V, Vodicka P and Soucek P. The role of ABC transporters in progression and clinical outcome of colorectal cancer. Mutagenesis 2012; 27: 187-196.
[36] Ohtsuki S, Kamoi M, Watanabe Y, Suzuki H, Hori S and Terasaki T. Correlation of induction of ATP binding cassette transporter A5 (ABCA5) and ABCB1 mRNAs with differentiation state of human colon tumor. Biol Pharm Bull 2007; 30: 1144-1146.
[37] Lai Wing Sun K, Correia JP and Kennedy TE. Netrins: versatile extracellular cues with diverse functions. Development 2011; 138: 2153-2169.
[38] Cirulli V and Yebra M. Netrins: beyond the brain. Nat Rev Mol Cell Biol 2007; 8: 296-306.
[39] Chaturvedi V, Fournier-Level A, Cooper HM and Murray MJ. Loss of Neogenin1 in human colorectal carcinoma cells causes a partial EMT and wound-healing response. Sci Rep 2019; 9: 4110.



Supplementary Figure 1. Exploratory analysis of transcriptomic data of cancerous and normal colon stem cells from Human Transcriptome Array 2.0 (HTA 2.0, Affymetrix). A. Principal Component Analysis (PCA) using normalized values of hybridization onto HTA 2.0 arrays of 3 independent samples of cancerous colon stem cells (HTA2_A, HTA2_B and HTA2_C, blue round symbols) and 3 independent samples of normal colon stem cells (HTA2_1, HTA2_2 and HTA2_3, magenta square symbols). PC1: principal component 1; PC2: principal component 2; VarExp: percentage of explained variance by the considered principal component. PCA analysis was performed using R (version 3.6.1) and *ggplot2* package for visualization. B. Heatmap of sample-to-sample distances using log-transformed normalized values of hybridization of cancerous and normal stem cells (tumor: blue; normal: magenta). R function *dist* was used to calculate Manhattan distance between samples and pheatmap package to assess overall similarity between samples. Sample-to-sample distance was assessed by a color gradient from yellow (lower) to red (largest). C. Hierarchical clustering of stem cells samples based on 162 genes differentially expressed between cancerous colon stem cells (HTA2_A, HTA2_B and HTA2_C) as compared to normal colon stem cells (HTA2_1, HTA2_2 and HTA2_3). Hierarchical clustering using Pearson's correlation coefficient and average linkage showed a clear separation of the two sets of cells.

Supplementary Table 1. Deregulated genes in colon cancer stem cells as compared to normal colon stem cells

A. Up-regulated genes in colon cancer stem cells identified on Human Transcriptome Array 2.0 (Affymetrix)

FAST DB® ID	Gene Symbol	Entrez Gene ID	Gene Coordinates (hg19)	Gene Name	Representative Transcript ID	Fold Change	*p* value
GSHG0041239	---	---	chr14 (+): 19596726-19596752	piRNA piR-37783	DQ599717	2.75	0.00262
GSHG0029102	PKIA	5569	chr8 (+): 79428336-79517502	protein kinase (cAMP-dependent, catalytic) inhibitor alpha	NM_181839	2.67	0.000698
GSHG0049942	LINC00613	100507528	chr4 (-): 136788138-136834835	long intergenic non-protein coding RNA 613	NR_103763	2.41	0.0312
GSHG0009996	NEO1	4756	chr15 (+): 73344825-73597547	neogenin 1	NM_002499	2.3	0.00148
GSHG0025500	HIST1H2AE	3012	chr6 (+): 26217148-26217711	histone cluster 1, H2ae	NM_021052	2.27	0.0412
GSHG0009955	SMAD3	4088	chr15 (+): 67358183-67487533	SMAD family member 3	NM_005902	2.18	0.0000234
GSHG0027256	AQP1	358	chr7 (+): 30737601-30965131	aquaporin 1 (Colton blood group)	NR_037598	2.16	0.00206
GSHG0026367	HIST1H3B	8358	chr6 (-): 26031817-26032290	histone cluster 1, H3b	NM_003537	2.09	0.00284
GSHG0026379	HIST1H3F	8968	chr6 (-): 26250361-26250868	histone cluster 1, H3f	BC096131	1.98	0.0364
GSHG0025491	HIST1H2AC	8334	chr6 (+): 26124366-26139336	histone cluster 1, H2ac	U90551	1.97	0.00862
GSHG0026402	HIST1H3J	8356	chr6 (-): 27858035-27858570	histone cluster 1, H3j	AB463735	1.93	0.0341
GSHG0025534	HIST1H2BO	8348	chr6 (+): 27861203-27861669	histone cluster 1, H2bo	NM_003527	1.91	0.0108
GSHG0025503	HIST1H2BH	8345	chr6 (+): 26251837-26253284	histone cluster 1, H2bh	AK310576	1.85	0.00916
GSHG0025488	HIST1H3C	8352	chr6 (+): 26045608-26047022	histone cluster 1, H3c	BC058834	1.83	0.0269
GSHG0023880	SLC1A3	6507	chr5 (+): 36606457-36688434	solute carrier family 1 (glial high affinity glutamate transporter), member 3	NM_004172	1.82	0.00862
GSHG0018777	PI3	5266	chr20 (+): 43803540-43805184	peptidase inhibitor 3, skin-derived	NM_002638	1.79	0.00214
GSHG0010450	ZNF280D	54816	chr15 (-): 56922376-57210776	zinc finger protein 280D	BC036541	1.79	0.0228
GSHG0025501	HIST1H3E	8353	chr6 (+): 26224427-26227699	histone cluster 1, H3e	BC052981	1.76	0.0432
GSHG0036837	MIR323A	442897	chr14 (+): 101492069-101492154	microRNA 323a	NR_029890	1.74	0.0338
GSHG0040746	MIR1245B	100616324	chr2 (-): 189842819-189842887	microRNA 1245b	NR_039947	1.73	0.000859
GSHG0024806	PLK2	10769	chr5 (-): 57749812-57756087	polo-like kinase 2	AF059617	1.73	0.00294
GSHG0021192	MRAS	22808	chr3 (+): 138066490-138124376	muscle RAS oncogene homolog	NM_012219	1.72	0.0012
GSHG0000214	CDA	978	chr1 (+): 20915441-20945398	cytidine deaminase	NM_001785	1.71	0.0191
GSHG0028372	SHFM1	7979	chr7 (-): 96318075-96339203	split hand/foot malformation (ectrodactyly) type 1	U41515	1.71	0.00618
GSHG0012878	CBX2	84733	chr17 (+): 77751962-77761449	chromobox homolog 2	NM_005189	1.69	0.00278
GSHG0021287	SMC4	10051	chr3 (+): 160117092-160152747	structural maintenance of chromosomes 4	NM_001288753	1.69	0.0199
GSHG0025285	CLTB	1212	chr5 (-): 175819457-175843570	clathrin, light chain B	NR_045724	1.68	0.00184
GSHG0012540	---	---	chr17 (+): 42023509-42027711	---	AK024231	1.68	0.000475
GSHG0026839	GSTM2P1	442245	chr6 (-): 111367622-111368757	glutathione S-transferase mu 2 (muscle) pseudogene 1	NR_002932	1.68	0.000516
GSHG0025505	HIST1H2BI	8346	chr6 (+): 26273144-26273640	histone cluster 1, H2bi	BC101655	1.68	0.0014
GSHG0007665	KITLG	4254	chr12 (-): 88886571-88974250	KIT ligand	NM_003994	1.68	0.000017
GSHG0025498	HIST1H2BF	8343	chr6 (+): 26199744-26200942	histone cluster 1, H2bf	BC056264	1.67	0.0341
GSHG0008423	SMAD9	4093	chr13 (-): 37418968-37494409	SMAD family member 9	NM_005905	1.67	0.0349
GSHG0031189	WDR34	89891	chr9 (-): 131395940-131419129	WD repeat domain 34	NM_052844	1.66	0.0025
GSHG0023987	CCNB1	891	chr5 (+): 68462837-68474068	cyclin B1	NM_031966	1.65	0.0352
GSHG0031506	RBM3	5935	chrX (+): 48432741-48439553	RNA binding motif (RNP1, RRM) protein 3	NM_006743	1.64	0.00226
GSHG0000780	GSTM5	2949	chr1 (+): 110254864-110260888	glutathione S-transferase mu 5	NM_000851	1.63	0.00244

GSHG0026398	HIST1H2AK	8330	chr6 (-): 27803287-27806153	histone cluster 1, H2ak	BC034487	1.62	0.018
GSHG0032562	ANXA8	653145	chr10 (+): 48255204-48271369	annexin A8	NM_001040084	1.61	0.00838
GSHG0002709	CREG1	8804	chr1 (-): 167510253-167525983	cellular repressor of E1A-stimulated genes 1	AK075375	1.61	0.0156
GSHG0020007	EIF3L	51386	chr22 (+): 38244875-38284789	eukaryotic translation initiation factor 3, subunit L	AK056129	1.6	0.033
GSHG0003019	HIST3H2A	92815	chr1 (-): 228644682-228645573	histone cluster 3, H2a	BC001193	1.6	0.0267
GSHG0029237	NDUFB9	4715	chr8 (+): 125551343-125580751	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9, 22 kDa	AK302286	1.6	0.0307
GSHG0000282	ARID1A	8289	chr1 (+): 27022518-27108601	AT rich interactive domain 1A (SWI-like)	NM_139135	1.58	0.00206
GSHG0009902	CCNB2	9133	chr15 (+): 59397284-59417249	cyclin B2	NM_004701	1.58	0.0302
GSHG0025521	HIST1H2AH	85235	chr6 (+): 27114861-27115343	histone cluster 1, H2ah	NM_080596	1.58	0.0163
GSHG0032726	H3C14	126961	chr1 (+): 149824181-149825836	H3 clustered histone 14	BC015544	1.58	0.0153
GSHG0029067	PREX2	80243	chr8 (+): 68864353-69143897	phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 2	AJ437636	1.58	0.00212
GSHG0005012	CNIH2	254263	chr11 (+): 66045672-66051685	cornichon family AMPA receptor auxiliary protein 2	NR_073079	1.57	0.0239
GSHG0021010	LINC00883	344595	chr3 (+): 106959539-107045811	long intergenic non-protein coding RNA 883	NR_028302	1.57	0.0421
GSHG0007975	GRK6P1	2871	chr13 (+): 21893231-21894745	G protein-coupled receptor kinase 6 pseudogene 1	AK295462	1.56	0.00158
GSHG0025530	HIST1H2BM	8342	chr6 (+): 27782822-27783267	histone cluster 1, H2bm	NM_003521	1.56	0.0382
GSHG0000656	LPHN2	23266	chr1 (+): 81771884-82458106	latrophilin 2	AK123422	1.56	0.00808
GSHG0016991	DLX1	1745	chr2 (+): 172950208-172954399	distal-less homeobox 1	NM_178120	1.55	0.0196
GSHG0026376	HIST1H2BG	8339	chr6 (-): 26215417-26216921	histone cluster 1, H2bg	BC082232	1.55	0.041
GSHG0030513	NCS1	23413	chr9 (+): 132934857-132999583	neuronal calcium sensor 1	NM_014286	1.55	0.0256
GSHG0017651	XPO1	7514	chr2 (-): 61705069-61765418	exportin 1	NM_003400	1.55	0.0241
GSHG0039160	MIR4522	100616277	chr17 (-): 25620936-25621022	microRNA 4522	NR_039748	1.54	0.011
GSHG0023435	PPP3CA	5530	chr4 (-): 101944587-102268634	protein phosphatase 3, catalytic subunit, alpha isozyme	EU192653	1.54	0.0019
GSHG0026396	HIST1H2AJ	8331	chr6 (-): 27782080-27782607	histone cluster 1, H2aj	BC133050	1.53	0.0151
GSHG0044888	LOC286297	286297	chr9 (-): 67017375-67032072	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1 like pseudogene	NR_046175	1.53	0.0193
GSHG0031064	LPAR1	1902	chr9 (-): 113635544-113801526	lysophosphatidic acid receptor 1	BC036034	1.53	0.0316
GSHG0000842	SLC22A15	55356	chr1 (+): 116519119-116612674	solute carrier family 22, member 15	NM_018420	1.52	0.00436
GSHG0006834	CCT2	10576	chr12 (+): 69979208-69995357	chaperonin containing TCP1, subunit 2 (beta)	NM_006431	1.51	0.00158
GSHG0007844	CIT	11113	chr12 (-): 120123595-120315095	citron (rho-interacting, serine/threonine kinase 21)/microRNA 1178	NM_007174	1.51	0.0042
GSHG0018819	CSE1L	1434	chr20 (+): 47662783-47713497	CSE1 chromosome segregation 1-like (yeast)	NR_045796	1.51	0.00588
GSHG0000779	GSTM1	2944	chr1 (+): 110230418-110236366	glutathione S-transferase mu 1	NM_146421	1.51	0.0135
GSHG0000778	GSTM2	2946	chr1 (+): 110210644-110252172	glutathione S-transferase mu 2 (muscle)	BC017836	1.51	0.00142
GSHG0011738	NUPR1	26471	chr16 (-): 28548662-28550495	nuclear protein, transcriptional regulator 1	NM_012385	1.51	0.0439
GSHG0021260	P2RY1	5028	chr3 (+): 152552731-152555845	purinergic receptor P2Y, G-protein coupled 1	NM_002563	1.51	0.0164
GSHG0000589	ROR1	4919	chr1 (+): 64239683-64647177	receptor tyrosine kinase-like orphan receptor 1	NM_005012	1.51	0.0129
GSHG0033969	VGLL3	389136	chr3 (-): 86987123-87040269	vestigial like 3 (Drosophila)	NM_016206	1.51	0.0198
GSHG0026856	HDAC2	3066	chr6 (-): 114257320-114292366	histone deacetylase 2	NR_033441	1.5	0.0273
GSHG0022084	MSL2	55167	chr3 (-): 135867760-135914688	male-specific lethal 2 homolog (Drosophila)	NM_018133	1.5	0.0244
GSHG0042985	LOC105375026	105375026	chr6 (+): 33871533-33871701	uncharacterized LOC105375026	DQ596554	1.5	0.00192
GSHG0044387	---	---	chr8 (-): 125934286-125934323	piRNA piR-58597	DQ591485	1.5	0.0427
GSHG0024992	STARD4	134429	chr5 (-): 110832645-110848292	StAR-related lipid transfer (START) domain containing 4	AK125317	1.5	0.0499

B. Down-regulated genes in colon cancer stem cells identified on Human Transcriptome Array 2.0 (Affymetrix)

FAST DB STABLE ID	Gene Symbol	Entrez Gene ID	Gene Coordinates (hg19)	Gene Name	Representative Transcript ID	Fold change	*P*-Value
GSHG0023383	MAPK10	5602	chr4 (-): 86936276-87374298	mitogen-activated protein kinase 10	BC051731	-3.44	0.0000194
GSHG0019238	SULF2	55959	chr20 (-): 46285656-46415360	sulfatase 2	NM_198596	-2.61	0.000544
GSHG0040996	MIR125B2	406912	chr21 (+): 17962557-17962645	microRNA 125b-2	NR_029694	-2.22	0.00464
GSHG0000833	OLFML3	56944	chr1 (+): 114522013-114524876	olfactomedin-like 3	NM_020190	-1.98	0.0000619
GSHG0042399	piR-35602	---	chr4 (-): 106407064-106407090	piRNA piR-35602	DQ597536	-1.96	0.00358
GSHG0048490	---	---	chr2 (+): 2875863-2876529	lnc-TRAPPC12-3	TCONS_00002797	-1.95	0.0177
GSHG0026526	UQCC2	84300	chr6 (-): 33664539-33679528	ubiquinol-cytochrome c reductase complex assembly factor 2	NM_032340	-1.91	0.0302
GSHG0043644	piR-46016	---	chr1 (+): 143962339-143962368	piRNA piR-46016	DQ577904	-1.9	0.0186
GSHG0013211	---	---	chr17 (-): 26700703-26701017	cDNA clone IMAGE: 40119874	BC128525	-1.85	0.00862
GSHG0007676	DCN	1634	chr12 (-): 91539036-91576806	decorin	NM_001920	-1.8	0.00232
GSHG0020772	LRRC2-AS1	83598	chr3 (+): 46598888-46601178	LRRC2 antisense RNA 1	NR_073385	-1.79	0.0231
GSHG0039530	piR-60625	---	chr18 (-): 11653988-11654031	piRNA piR-60625	DQ594513	-1.78	0.02
GSHG0047286	---	---	chr12 (+): 111376095-111396124	lnc-CCDC63-2	TCONS_00020561	-1.77	0.0153
GSHG0037870	piR-52451	---	chr15 (-): 51592525-51592595	piRNA piR-52451	DQ585339	-1.77	0.00364
GSHG0032370	FGF13	2258	chrX (-): 137713735-138287216	fibroblast growth factor 13	NM_001139501	-1.76	0.0015
GSHG0050901	---	---	chr7 (-): 50518216-50521151	lnc-FIGNL1-1	TCONS_00013800	-1.72	0.0183
GSHG0040541	MIR216B	100126319	chr2 (-): 56227849-56227930	microRNA 216b	NR_030623	-1.72	0.0151
GSHG0040506	MIR4429	100616469	chr2 (-): 11680731-11680803	microRNA 4429	NR_039627	-1.72	0.0382
GSHG0028286	PMS2P3	5387	chr7 (-): 75137070-75157453	postmeiotic segregation increased 2 pseudogene 3	NR_028059	-1.72	0.031
GSHG0044676	piR-53295	---	chr5 (-): 43495263-43495292	piRNA piR-53295	DQ586183	-1.71	0.0266
GSHG0016845	RAB6C	84084	chr2 (+): 130737612-130740313	RAB6C, member RAS oncogene family	AK055504	-1.71	0.002
GSHG0011105	LOC101928595	101928595	chr16 (+): 30107751-30116777	uncharacterized LOC101928595	AK095480	-1.7	0.0327
GSHG0010984	NDE1	54820	chr16 (+): 15793016-15795769	nudE neurodevelopment protein 1	AK123247	-1.7	0.00234
GSHG0043002	piR-50599	---	chr6 (+): 33873327-33873383	piRNA piR-50599	DQ583487	-1.69	0.00724
GSHG0022022	---	---	chr3 (-): 126167060-126169613	Sequence 1836 from Patent EP1308459	AX748311	-1.69	0.000136
GSHG0043468	---	---	chr6 (-): 142334792-142334865	Sequence 59 from Patent EP2374884	JA611295	-1.68	0.0279
GSHG0026802	ASCC3	10973	chr6 (-): 100956090-101329248	activating signal cointegrator 1 complex subunit 3	NM_006828	-1.66	0.0159
GSHG0051189	---	---	chr8 (-): 124009622-124012419	lnc-DERL1-1	TCONS_00015156	-1.66	0.0044
GSHG0042019	SCARNA22	677770	chr4 (+): 1976363-1976487	small Cajal body-specific RNA 22	NR_003004	-1.66	0.00108
GSHG0015336	JSRP1	126306	chr19 (-): 2252250-2256422	junctional sarcoplasmic reticulum protein 1	NM_144616	-1.64	0.00508
GSHG0037727	MIR4716	100616332	chr15 (-): 49461267-49461350	microRNA 4716	NR_039866	-1.64	0.0196
GSHG0011206	---	---	chr16 (+): 56334528-56336816	---	AK093105	-1.63	0.00172
GSHG0024284	IGIP	492311	chr5 (+): 139505517-139508977	IgA-inducing protein	NM_001007189	-1.63	0.047
GSHG0037845	piR-38299	---	chr15 (-): 51585622-51585716	piRNA piR-38299	DQ600233	-1.63	0.0053
GSHG0013552	SPOP	8405	chr17 (-): 47676248-47755525	speckle-type POZ protein	NM_003563	-1.63	0.000697
GSHG0032293	SLC25A5-AS1	100303728	chrX (-): 118599997-118603061	SLC25A5 antisense RNA 1	BC028211	-1.62	0.0312
GSHG0020060	TEF	7008	chr22 (+): 41763337-41795332	thyrotrophic embryonic factor	NM_001145398	-1.62	0.00456
GSHG0020003	LGALS1	3956	chr22 (+): 38071613-38075813	lectin, galactoside-binding, soluble 1	NM_002305	-1.61	0.0216

GSHG0028308	MAGI2	9863	chr7 (-): 77646374-79082890	membrane associated guanylate kinase, WW and PDZ domain containing 2	NM_012301	-1.61	0.0237
GSHG0026563	ETV7	51513	chr6 (-): 36321998-36356172	ets variant 7	NM_001207039	-1.6	0.00504
GSHG0016824	INHBB	3625	chr2 (+): 121103719-121109383	inhibin, beta B	NM_002193	-1.6	0.0123
GSHG0048626	---	---	chr2 (+): 123495841-123506009	lnc-TSN-8	TCONS_00003846	-1.6	0.0298
GSHG0031382	TMSB4X	7114	chrX (+): 12993222-12995345	thymosin beta 4, X-linked	NM_021109	-1.6	0.00392
GSHG0019257	TMSB4XP6	7120	chr20 (-): 49457129-49457312	thymosin beta 4, X-linked pseudogene 6	BC112282	-1.6	0.0105
GSHG0041410	piR-58863	---	chr22 (-): 37756430-37756548	piRNA piR-5886	DQ591751	-1.59	0.0464
GSHG0000168	SLC25A34	284723	chr1 (+): 16062809-16067885	solute carrier family 25, member 34	NM_207348	-1.59	0.0015
GSHG0033316	ZNF580	51157	chr19 (+): 56152301-56154835	zinc finger protein 580	AL359054	-1.59	0.0188
GSHG0027496	ADAM22	53616	chr7 (+): 87563566-87832204	ADAM metallopeptidase domain 22	NM_021723	-1.58	0.0149
GSHG0031904	---	---	chrX (+): 152864793-152865337	cDNA clone IMAGE: 4797878	BC030106	-1.58	0.0184
GSHG0051071	---	---	chr8 (+): 102300022-102305710	lnc-GRHL2-3	TCONS_00014812	-1.58	0.0409
GSHG0028752	MNX1	3110	chr7 (-): 156786745-156803347	motor neuron and pancreas homeobox 1	AY927465	-1.58	0.0131
GSHG0012991	OR1E2	8388	chr17 (-): 3335902-3337145	olfactory receptor, family 1, subfamily E, member 2	AB529302	-1.58	0.00464
GSHG0039820	SNORD35B	84546	chr19 (+): 50000976-50001063	small nucleolar RNA, C/D box 35B	NR_001285	-1.58	0.0258
GSHG0033194	BRI3BP	140707	chr12 (+): 125478194-125511045	BRI3 binding protein	AF284094	-1.57	0.0252
GSHG0038507	piR-51963	---	chr15 (+): 62543174-62543240	piRNA piR-51963	DQ584851	-1.57	0.0227
GSHG0034434	SNORD85	692200	chr1 (-): 31441010-31441084	small nucleolar RNA, C/D box 85	NR_003066	-1.57	0.00304
GSHG0007576	AGAP2	116986	chr12 (-): 58118076-58135944	ArfGAP with GTPase domain, ankyrin repeat and PH domain 2	NM_014770	-1.56	0.000297
GSHG0049182	---	---	chr21 (+): 40739091-40742922	lnc-WRB-1	TCONS_00029021	-1.56	0.00628
GSHG0027541	BHLHA15	168620	chr7 (+): 97840778-97844752	basic helix-loop-helix family, member a15	BX648200	-1.55	0.0129
GSHG0008949	---	---	chr14 (+): 71276914-71280446	cDNA FLJ39181 fis, clone OCBBF2004235	AK096500	-1.55	0.0093
GSHG0010765	LINS	55180	chr15 (-): 101109428-101142445	lines homolog (Drosophila)	NM_001040616	-1.55	0.0328
GSHG0001178	PRRX1	5396	chr1 (+): 170632303-170708541	paired related homeobox 1	NM_022716	-1.55	0.00168
GSHG0006004	RAB6A	5870	chr11 (-): 73386683-73472201	RAB6A, member RAS oncogene family	NM_001243718	-1.55	0.00188
GSHG0022963	TRIM60	166655	chr4 (+): 165953150-165962896	tripartite motif containing 60	AX747987	-1.55	0.0196
GSHG0018008	WTH3DI	150786	chr2 (-): 132118060-132121731	RAB6C-like	NM_001077637	-1.55	0.0449
GSHG0024672	ANKH	56172	chr5 (-): 14704910-14871887	ANKH inorganic pyrophosphate transport regulator	NM_054027	-1.54	0.000583
GSHG0026372	HIST1H1T	3010	chr6 (-): 26107640-26108364	histone cluster 1, H1t	NM_005323	-1.54	0.0388
GSHG0050316	---	---	chr5 (-): 116156689-116165883	lnc-SEMA6A-4	TCONS_00009758	-1.54	0.0124
GSHG0034139	MIR30E	407034	chr1 (+): 41220027-41220118	microRNA 30e	NR_029846	-1.54	0.0358
GSHG0013232	ABHD15	116236	chr17 (-): 27887691-27894048	abhydrolase domain containing 15	NM_198147	-1.53	0.0208
GSHG0001051	APOA1BP	128240	chr1 (+): 156561548-156566601	apolipoprotein A-I binding protein	AK294835	-1.53	0.00268
GSHG0004058	ARHGAP22	58504	chr10 (-): 49654068-49864310	Rho GTPase activating protein 22	NR_045675	-1.53	0.0166
GSHG0031828	DDX26B	203522	chrX (+): 134654547-134716462	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 26B	NM_182540	-1.53	0.0000181
GSHG0003400	ADO	84890	chr10 (+): 64564513-64568238	2-aminoethanethiol (cysteamine) dioxygenase	NM_032804	-1.52	0.0371
GSHG0001773	CHCHD2P6	645317	chr1 (-): 15912631-15930121	coiled-coil-helix-coiled-coil-helix domain containing 2 pseudogene 6	AK091499	-1.52	0.0318
GSHG0047720	---	---	chr14 (+): 102097395-102100579	lnc-DIO3.1-1	TCONS_00022639	-1.52	0.0339
GSHG0039636	MIR3591	100616357	chr18 (-): 56118312-56118384	microRNA 3591	NR_039899	-1.52	0.00848
GSHG0016189	UBE2M	9040	chr19 (-): 59067080-59070343	ubiquitin-conjugating enzyme E2M	NM_003969	-1.52	0.0267
GSHG0016505	AFTPH	54812	chr2 (+): 64751439-64820136	aftiphilin	NM_203437	-1.51	0.00438

GSHG0019631	KRTAP19-5	337972	chr21 (-): 31873975-31874435	keratin associated protein 19-5	AB096946	-1.51	0.0199
GSHG0038690	LOC730183	730183	chr16 (-): 30709025-30709810	uncharacterized LOC730183	NM_001256932	-1.51	0.0186
GSHG0002075	TAL1	6886	chr1 (-): 47681962-47697892	T-cell acute lymphocytic leukemia 1	NM_001287347	-1.51	0.00196
GSHG0031363	VCX3B	425054	chrX (+): 8432871-8434551	variable charge, X-linked 3B	NM_001001888	-1.51	0.0137
GSHG0012550	FZD2	2535	chr17 (+): 42634812-42638629	frizzled class receptor 2	NM_001466	-1.5	0.0233
GSHG0049599	---	---	chr3 (-): 129993880-129995570	lnc-TMCC1-4	TCONS_00006657	-1.5	0.00364

Supplementary Table 2. Functional enrichment analysis of differentially expressed genes in cancerous colon stem cells as compared to normal colon stem cells

A. Top-enriched Gene Ontology (GO) biological process analysis

GO ID	GO Name	Differential gene counts in GO	Gene amount in GO	Differential genes in GO	Enrichment score	*p*. value	p.adjusted (Benjamini-Hochberg correction)
GO:0006334	nucleosome assembly	12	119	HIST1H2BO, HIST1H3J, HIST1H2BM, HIST1H1T, HIST1H2BF, HIST1H2BG, HIST1H2BI, HIST1H2BH, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	16.9	1.06E-10	7.63E-8
GO:0060968	regulation of gene silencing	5	11	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	76.3	3.63E-7	1.31E-4
GO:0006342	chromatin silencing	6	45	HIST1H2AC, HIST1H2AE, HIST1H2AH, HIST3H2A, HIST1H2AK, HIST1H2AJ	22.4	6.52E-6	0.001
GO:0032200	telomere organization	5	27	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	31.1	1.8E-5	0.003
GO:0006335	DNA replication-dependent nucleosome assembly	5	32	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	26.2	3.6E-5	0.005
GO:0000183	chromatin silencing at rDNA	5	37	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	22.7	6.46E-5	0.008
GO:0051290	protein heterotetramerization	5	42	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	20	1.07E-4	0.011
GO:0045814	negative regulation of gene expression, epigenetic	5	50	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	16.8	2.13E-4	0.019
GO:0045815	positive regulation of gene expression, epigenetic	5	62	HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	13.5	4.88E-4	0.038
GO:000961	response to mechanical stimulus	4	59	CCNB1, INHBB, P2RY1, DCN	11.4	0.005	0.26
GO:0060395	SMAD protein signal transduction	4	62	INHBB, SMAD9, MAGI2, SMAD3	10.8	0.006	0.26
GO:1901687	glutathione derivative biosynthetic process	3	22	GSTM1, GSTM2, GSTM5	22.9	0.007	0.3
GO:0042493	response to drug	7	304	CCNB1, XPO1, SLC1A3, HDAC2, LGALS1, AQP1, RAB6C	3.9	0.009	0.34

B. Top-enriched KEGG pathway analysis

Pathway ID	Pathway Name	Differential gene counts in pathway	Gene amount in pathway	Differential genes in pathway	Enrichment score	*p*. value	p.adjusted (Benjamini-Hochberg correction)
hsa05034	Alcoholism	19	177	HIST1H2AC, HIST1H3J, HIST1H2BF, HIST1H2BG, HIST1H2BH, HIST1H2AE, PKIA, HIST1H2BO, HIST1H2BM, HDAC2, HIST1H2BI, HIST1H3B, HIST1H3C, HIST1H2AH, HIST1H2AK, HIST3H2A, HIST1H2AJ, HIST1H3E, HIST1H3F	12.9	8.87E-16	1.0E-13
hsa05322	Systemic lupus erythematosus	17	134	HIST1H2AC, HIST1H3J, HIST1H2BF, HIST1H2BG, HIST1H2AE, HIST1H2BH, HIST1H2BO, HIST1H2BM, HIST1H2BI, HIST1H3B, HIST1H3C, HIST1H2AH, HIST1H2AK, HIST3H2A, HIST1H2AJ, HIST1H3E, HIST1H3F	15.3	3.68E-15	2.07E-13
hsa05202	Transcriptional misregulation in cancer	8	167	HIST1H3J, ETV7, HDAC2, NUPR1, HIST1H3B, HIST1H3C, HIST1H3E, HIST1H3F	5.8	3.73E-4	0.014
hsa04068	FoxO signaling pathway	6	134	CCNB1, CCNB2, PLK2, SMAD3, MAPK10, AGAP2	5.4	0.0045	0.119

Summary of top-enriched biological processes or pathways performed with DAVID (Database for Annotation, Visualization and Integrated Discovery) database (version 6.8) using Gene Ontology (GO) biological process (A) and KEGG pathway analysis (B).