Am J Cancer Res 2022;12(1):371-380
www.ajcr.us /ISSN:2156-6976/ajcr0138973

*Original Article*

# Analyzing large scale gene expression data in colorectal cancer reveals important clues; CLCA1 and SELENBP1 downregulated in CRC not in normal and not in adenoma

Fariborz Asghari Alashti[1,2], Bahram Goliaei[1], Zarrin Minuchehr[3]

*[1]Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran; [2]Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada; [3]National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran*

Received September 8, 2021; Accepted November 26, 2021; Epub January 15, 2022; Published January 30, 2022

**Abstract:** Early detection of colorectal cancer (CRC) increases the chances of survival and reduces the therapeutic problems and costs of treatment. Since molecular biomarkers can help us diagnose colorectal cancer early, we need to identify novel gene for predicting the early stages of tumorigenesis. Here, we integrated five independent CRC gene expression datasets derived from expression profiling by array comparing CRC with normal samples in: GSE21510, GSE4107, GSE25071, GSE15781 dataset, and GSE8671 dataset, including 64 samples from 32 patients comparing 32 colonic normal mucosa with 32 colorectal adenoma. To detect genes that expressed differentially in experimental circumstances of these datasets, we used web tool of GEO2R to compare groups of samples in the GEO data series. Furthermore, we constructed the protein-protein interactions network by STRING database for mostly downregulated genes and the expression of their members in PPI network were studied into five datasets separately. Also, the level of expression of selected biomarker genes in different stages of CRC compared to normal was studied. Our data revealed 17 common downregulated genes (average fold change (FC) in five tests ≥6) in CRC in comparison with normal (Test 1 to Test 4) and in adenoma compared with normal (Test 5). Studying of gene expression of PPI network members of these downregulated genes led to identifying of CLCA1, SELENBP1, CWC25, ACOT11, GUCY2C and ALDH1A1 as suppressor genes and PTGS2, PROCR, MOCS3 and NFS1 as oncogenes which respectively downregulated and upregulated in CRC. Since decreasing of gene expression was seen in CRC comparing with normal and due to no different expression seen for these 10 genes in adenoma, they, especially CLCA1 and SELENBP1, could be considered as biomarkers for early detection of CRC. Before using these signature genes in the clinic; however, further validations are required.

**Keywords:** Colorectal cancer, GEO datasets, protein-protein interactions, STRING, biomarker, early detection

## Introduction

Colorectal cancer (CRC) is one of the most common malignancies and the second most common cause of cancer death in the United States when women and men are united. During 2008 through 2017, annually CRC death rates dropped by 3% in persons aged 65 years and older and this rate increased by 1.3% in people younger than 50 years. Although CRC is still one of the chief health difficulties in worldwide, due to early detection, information of risk factors, and prevention, death from it has been decreased [1-3].

Genes, mRNA transcripts, proteins and other variables which are considered as molecular signatures can be used as biomarkers for a particular phenotype of cells and tissues [4]. In CRC, by using gene expression profiling, alterations at the transcriptional level are identified and can be considered as early biomarker which help to detect disease or to find therapeutic solutions [5]. Many differentially expressed genes, identified as molecular signatures, have been recognized by techniques of RNA-seq, microarrays or qPCR and used for early detection of CRC [6, 7]. The Gene Expression Omnibus (GEO) at the National Center for

Biotechnology Information (NCBI) is a database which keep molecular abundance data produced by an extensive variety of high-throughput measuring techniques including genomic DNA, microarray-based experiments, and protein molecules, mass spectrometry proteomic technology and serial analysis of gene expression (SAGE) [8]. Gene expression signatures from the Gene Expression Omnibus (GEO) is very valuable in which signatures have been applied for suggesting new drugs against cancer [9].

Proteins seldom act unaccompanied because their functions incline to be regulated. Frequent protein components which build molecular machines carry out many molecular processes within a cell. When two or more protein molecules are in high specificity physical contacts, biochemical events happened. Information of signal transduction, quantum chemistry, biochemistry and molecular dynamics of these types of contacts assists the making of PPI networks. PPI is vital because helps to the understanding of cell physiology in normal and disease conditions [10, 11]. For the study of PPI, posttranslational modifications, homologous pairs, intracellular localization, phylogenetic profiling and identifying structural patterns are considered [12]. STRING, a known biological database, predicts protein-protein interactions according to the information of genomic context predictions, high-throughput lab experiments, co-expression, automated textmining and previous knowledge in databases [13].

In this study, we united five independent CRC gene expression datasets from GEO database, which led to the detection of ten genes as signature related to CRC which can help us the recognition of CRC in the early stage.

## Materials and methods

### *Patient information*

The present study was carried out on five different CRC cohorts from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI):

(1) The GSE21510 dataset [14], this experiment was done for the purpose of identifying a novel biomarker for CRC by microarray analysis of the laser microdissection and oligonucleotide on 104 patients with CRC and 77 normal samples. The comparison of gene expression between CRC and normal carried out and in this paper titled Test 1.

(2) The GSE4107 dataset [15], by the aim of searching genes expressed differentially in early onset CRC was performed by using DNA chip technology. In this study RNA from colonic mucosa of healthy controls and patients were analyzed. All of samples including 12 samples of CRC patients and 10 samples of healthy controls were collected from people in the age of less than 50 years old. This experiment titled Test 2 in this paper.

(3) The GSE25071 dataset [16], due to an increasing of CRC with age, this experiment analyzed the genome-wide gene expression levels from at an early age and at higher age patients diagnosed CRC; including 41 CRC patients in two group of 24 young age samples (mean, 43; range, 28-53 years) and 17 old age samples (mean, 79 years; range, 69-87 years) and 9 normal samples. For our analysis, we considered all CRC patients in one group and compared them with all normal samples and titled it Test 3 in this paper.

(4) The GSE15781 dataset [17]; the aim of this study was to recognize the properties of preoperative radiochemotherapy (PRT) on gene expression before and after PRT in tumour and normal colon rectal tissue from the same patients, including 42 samples (13 tumor tissues, non-irradiated; 10 normal tissues, non-irradiated; 9 tumor tissues, irradiated and 10 normal tissues, irradiated) collected from ten patients. For our study we just measured non-irradiated samples and titled this experiment Test 4 in this paper.

(5) The GSE8671 dataset [18], because it is believed that Colorectal cancers rise mostly from adenomas, this experiment was accomplished to examine differentiation of gene expression in CRC and adenoma, including 64 samples from 32 patients (32 colonic normal mucosa and 32 colorectal adenoma). We titled this experiment Test 5 in this paper. Although we used this test to compare its data with other four tests to recognize common regulated genes, we also deliberate this test as a control to compare the results of this test on adenoma with other four tests on CRC.

*Microarray data analysis*

The GSE21510, GSE4107, GSE25071, GSE15781 and GSE8671 raw gene expression datasets were retrieved from the GEO and were loaded into GEO2R tool to investigate these data series. GEO2R is an interactive web tool which permits users to compare two or more groups of samples in a GEO series to recognize genes that are expressed under different experimental situations. After analyzing, the results are accessible as a table of genes arranged in order of importance.

Also, since the results of microarray data analysis of different CRC stages were collected in GSE21510 [14], in this study, this database was analyzed to evaluate the expression level of selected biomarker genes in different CRC stages. Four studies were performed separately. In the first study, 4 normal samples in the first stage with 15 cancer samples of this stage, the second study, 8 normal samples in the second stage with 46 cancer samples of this stage, the third study, 8 normal samples in the third stage with 39 cancer samples of this stage and the fourth study, 4 normal samples in the fourth stage with 23 cancer samples in this stage were deliberated.

*Protein-protein interaction*

Because of the importance of PPI in forming of macromolecular structures and activity of enzymes that are vital to almost all cellular processes, we used STRING [13] to make PPI network for selected genes.

## Results

*The comparison of gene expression level in 5 data series by GEO2R*

To create a gene expression panel associated with CRC, we analyzed five independent CRC gene expression datasets (GSE21510, GSE4107, GSE25071, GSE15781 and GSE8671) and recognized the differentially expressed transcripts between patient cells and normal samples. To identify suppressor genes associated with CRC, we considered the downregulated genes in the result of analysis of GSE21510, GSE4107, GSE25071, GSE15781 and GSE8671. For this purpose we just deliberated data which were downregulated in all 5 tests. We carefully examined the results in all 5 experiments and calculated their average LogFC separately for each gene in these 5 experiments and identified the genes that had a LogFC of -1.5 or less in all experiments with their average as genes with the lowest gene expression compared to the control samples and selected them as vital genes in CRC. The 21 most downregulated expression genes with the average of less than -6 FC were revealed (**Table 1**).

*Prediction of protein-protein interactions*

Based on the importance of PPI in cellular activities, in order to know what other genes may be affected by reduction of gene expression, the PPI network for the genes in **Table 1** were constructed by STRING (version 11.5) (**Figure 1**). Firstly, it was done for chloride channel accessory 4 (CLCA4), the most downregulated gene expression in **Table 1**. There are 21 nodes in this network. The interactions might contain physical and functional relations although joint to a shared function does not essentially mean they physically bind to each other.

*Evaluation of the expression of genes in the PPI network of CLCA4 gene*

Analysis of the expression of 21 genes in the PPI network related to the CLCA4 gene in **Figure 1** shows that some of them have a certain trend in their expression in cancer cells and adenoma compared to normal cells, which can be very important for the diagnosis and treatment of this disease. Because colorectal cancers are believed to be raised primarily by adenomas, differences in the progression from adenoma to cancer can help us learn more about this and ultimately treat it. Therefore, the study of changes in gene expression in these two stages will be very important. The results of the study in the **Table 2** show that although some genes in the study groups of cancer cells were reduced or increased compared to the control samples, there was no differences between adenoma with control, which can be named as influential genes in the transition phase. There are genes that show a decrease or increase in gene expression in both cancer groups and adenoma groups compared to normal samples, which can be considered as primary cancer-providing genes. According to STRING, type of

**Table 1.** Downregulated genes in 5 datasets

| Gene Symbol | LogFC | | | | | | Fold Change (FC) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average |
| CLCA4 | -7.9 | -1.7 | -7.1 | -4.3 | -5.3 | -5.2 | -241 | -3 | -138 | -19 | -38 | -88 |
| AQP8 | -7.1 | -2.3 | -6.7 | -5.9 | -6.5 | -5.5 | -141 | -5 | -102 | -58 | -90 | -79 |
| CA1 | -6.9 | -1.6 | -4.6 | -4.6 | -5.5 | -4.4 | -80 | -3 | -179 | -35 | -7 | -61 |
| CA4 | -6.4 | -1.7 | -4.7 | -3.8 | -3.3 | -4.1 | -120 | -3 | -25 | -24 | -45 | -43 |
| GUCA2A | -6.3 | -1.5 | -7.5 | -5.1 | -2.9 | -5.1 | -37 | -3 | -60 | -32 | -73 | -41 |
| CA2 | -6.3 | -1.9 | -5.0 | -4.1 | -3.6 | -4.3 | -58 | -4 | -75 | -20 | -9 | -33 |
| CD177 | -6.2 | -2.4 | -4.2 | -2.8 | -3.6 | -3.9 | -35 | -5 | -68 | -15 | -39 | -32 |
| ZG16 | -5.9 | -1.9 | -6.2 | -4.3 | -3.2 | -4.6 | -77 | -4 | -31 | -17 | -12 | -28 |
| SLC4A4 | -5.4 | -2.0 | -3.5 | -2.9 | -2.1 | -3.2 | -84 | -3 | -27 | -14 | -10 | -27 |
| CEACAM7 | -5.2 | -1.6 | -5.9 | -5.0 | -6.2 | -4.4 | -73 | -5 | -19 | -7 | -12 | -23 |
| ADH1C | -5.1 | -2.2 | -6.1 | -3.9 | -5.3 | -4.3 | -16 | -7 | -38 | -10 | -6 | -15 |
| DHRS9 | -4.8 | -2.0 | -3.5 | -2.9 | -2.1 | -3.0 | -42 | -4 | -11 | -7 | -4 | -14 |
| PKIB | -4.8 | -2.4 | -4.2 | -2.8 | -3.6 | -3.6 | -30 | -3 | -16 | -10 | -7 | -13 |
| CHP2 | -4.7 | -1.8 | -3.6 | -2.1 | -2.0 | -2.8 | -19 | -4 | -19 | -11 | -10 | -13 |
| CPM | -4.3 | -2.0 | -4.2 | -3.5 | -3.3 | -3.5 | -26 | -3 | -12 | -4 | -4 | -10 |
| HSD17B2 | -4.0 | -2.8 | -5.2 | -3.4 | -2.5 | -3.9 | -23 | -5 | -5 | -4 | -5 | -8 |
| HPGD | -3.5 | -1.5 | -2.3 | -2.5 | -2.4 | -2.5 | -11 | -3 | -5 | -6 | -5 | -6 |

The result of GEO2R analysis on the raw data of 5 experiments in the form of genes with the highest reduction in gene expression in CRC compared with normal (Test 1 to Test 4) and adenoma with normal (Test 5). The obtained LogFC were converted to FC with the FC=2^logFC formula, minus indicates a decrease in gene expression in CRC patients compared to control samples.

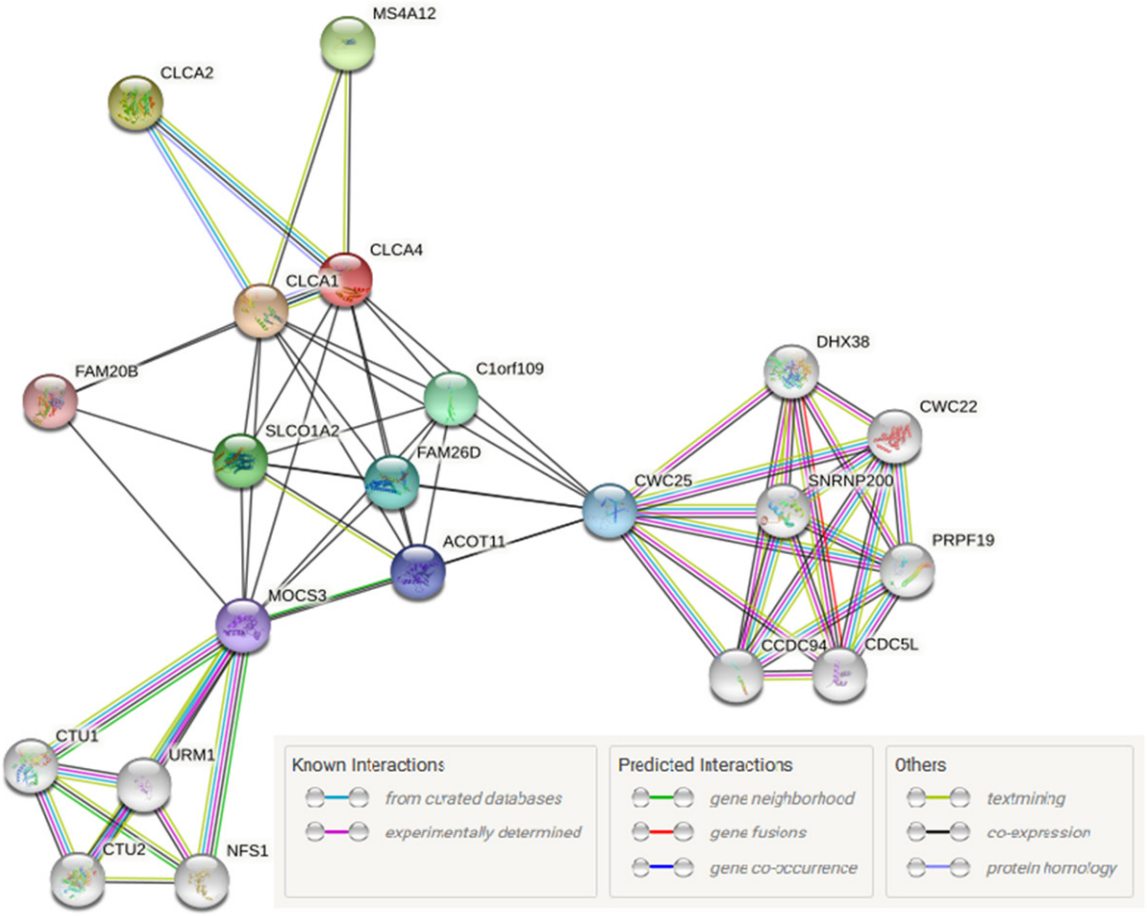

**Figure 1.** The PPI STRING network and the associated data for CLCA4 input (updated to version 11.5).

**Table 2.** Associated genes in STRING PPI network for CLCA4

| Gene Symbol | LogFC | | | | | | Fold Change (FC) | | | | | | Comment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average | |
| CLCA1 | -4.0 | -2.4 | -5.1 | -2.6 | | -3.5 | -16.2 | -5.3 | -35.2 | -6.1 | | -15.7 | A |
| CWC25 | -0.2 | -0.03 | -3.2 | -0.05 | | -0.9 | -1.2 | -1.02 | -9.3 | -1.03 | | -3.1 | |
| ACOT11 | -1.5 | -1.0 | -0.5 | -0.5 | | -0.9 | -2.7 | -2.1 | -1.4 | -1.4 | | -1.5 | |
| MOCS3 | 0.1 | 0.6 | 1.5 | 0.8 | | 0.8 | 1.1 | 1.5 | 2.9 | 1.7 | | 1.8 | B |
| NFS1 | 0.8 | 0.09 | 0.3 | 0.35 | | 0.4 | 1.7 | 1.1 | 1.3 | 1.3 | | 1.1 | |
| MS4A12 | -7.1 | -1.5 | -6.4 | -4.6 | -5.9 | -5.1 | -139 | -3 | -82 | -24 | -59 | -61 | C |
| C1orf109 | 1.9 | 0.4 | 1.0 | 0.8 | 0.5 | 0.9 | 3.7 | 1.4 | 1.9 | 1.8 | 1.4 | 2.1 | D |
| CWC22 | 1.1 | 0.3 | 0.4 | 0.3 | 0.3 | 0.5 | 2.1 | 1.2 | 1.4 | 1.2 | 1.2 | 1.4 | |
| SNRNP200 | 0.8 | 0.1 | 0.6 | 0.6 | 0.3 | 0.5 | 1.7 | 1.1 | 1.5 | 1.5 | 1.3 | 1.4 | |
| PRPF19 | 0.6 | 0.1 | 0.6 | 1.1 | 0.8 | 0.6 | 1.5 | 1.1 | 1.5 | 2.2 | 1.7 | 1.6 | |
| CDC5L | 0.1 | 0.02 | 0.5 | 0.3 | 0.4 | 0.3 | 1.1 | 1.0 | 1.4 | 1.2 | 1.3 | 1.2 | |
| CLCA2 | -0.3 | -1.1 | | -0.4 | | -0.6 | -1.3 | -2.1 | | -1.4 | | -1.6 | E |
| FAM26D | 0.0 | 0.7 | | 0.1 | | 0.3 | 1.0 | 1.6 | | 1.1 | | 1.2 | |
| SLCO1A2 | -0.6 | 0.04 | | -0.03 | | | -1.5 | 1.0 | | -1.0 | | | |
| FAM20B | 0.1 | 0.2 | -0.03 | 0.2 | | | 1.0 | 1.2 | -1.0 | 1.1 | | | |
| CTU1 | | | | 0.3 | | | | | | 1.3 | | | |
| CTU2 | -0.3 | 0.5 | | | 0.9 | | -1.3 | 1.4 | | | 1.8 | | |
| URM1 | -0.3 | 0.6 | 1.5 | 0.2 | 0.3 | | -1.2 | 1.5 | 2.8 | 1.2 | 1.3 | | |
| DHX38 | -0.8 | 0.3 | 0.4 | -0.1 | | | -1.7 | 1.2 | 1.4 | -1.1 | | | |
| CCDC94 | -0.4 | 0.1 | 1.2 | 0.1 | | | -1.3 | 1.1 | 2.3 | 1.0 | | | |

The results of the study of the expression of genes related to proteins in the PPI network related to CLCA4. A: The genes in this group showed decreased expression in experiments comparing cancer groups with normal cells (Test 1 to Test 4), while in experiments comparing the adenoma group with normal cells (Test 5), no difference was seen; B: The genes in this group showed increased expression in experiments comparing cancer groups with normal cells, while in experiments comparing the adenoma group with normal cells, no difference was seen; C: The genes of this group showed decreased expression in experiments related to comparing cancer groups with normal cells and in experiments related to comparing adenoma groups with normal cells; D: The genes of this group have shown increased expression in experiments related to comparing cancer groups with normal cells and in experiments related to comparing adenoma groups with normal cells; E: Changes in the expression of genes in this group in experiments related to the comparison of cancer groups with normal cells and in experiments related to the comparison of adenoma groups with normal cells do not show a specific trend. In some experiments, there is an increase and in some cases a decrease and sometimes no change between them.

PPI connection between, chloride channel accessory 1 (CLCA1) and membrane spanning 4-domains A12 (MS4A12) with CLCA4 in PPI network of genes in **Table 1** is co-expression and for other genes does not show this kind of connection.

*Construction of the PPI network for each gene in **Table 1***

For each of the genes in **Table 1**, using STRING, we predicted the PPI network separately and the names of the members are shown in **Table 3**. Then, we surveyed the gene expression of each protein in the respective PPI network in the 5 studied datasets. Some of the genes that show a decrease or increase in expression in all Tests 1 to 5, or show the same trend in all Tests 1 to 4, but do not show a difference in their expression in Test 5. The results are shown in **Table 4**.

Among the genes related to the PPI (**Table 3**) of each gene in **Table 1**, we just considered which had decreased or increased expression in all 5 Tests, or decreased or increased expression in four tests of 1-4 comparing normal cells with cancer cells, but in Test 5, which compared normal cells with adenoma is no change in their expression (**Table 4**). Other genes within the network (**Table 3**) which do not show a clear and uniform trend in increasing or decreasing gene expression were not mentioned. The association of genes in the PPI communication network with the connectivity type of co-expression has interestingly had a similar trend in gene expression to the related gene.

**Table 3.** Associated genes in the PPI network of each gene in **Table 1** obtained from the STRING based on the rank of physical and/or functional connectivity to the input gene

| Genes from **Table 1** | PPI network genes | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| AQP8 | AQP10 | AQP11 | AQP12A | AQP7 | AQP3 | AQP9 | AQP1 | LCMT1 | MIP | ARHGAP17 |
| CA1 | STAT4 | CYP24A1 | AHSP | EPB42 | SLC4A1 | ALAS2 | HBD | GYPB | SELENBP1 | SFMBT1 |
| GUCA2A | GUCY2C | NPR3 | TMIGD1 | EMD | C7orf49 | PDZD3 | GUCA2B | STH | SLC26A3 | CLCA1 |
| CA2 | SLC9A1 | SLC4A4 | SLC4A1 | HSPD1 | CDH1 | CTNNB1 | CTNND1 | RAP1B | CTNNA1 | RAP1A |
| CD177 | PECAM1 | PRTN3 | CNTN5 | ITGAM | CNTN6 | KCNIP3 | PLAUR | FAM101B | LYPD3 | PROCR |
| ZG16 | HORMAD1 | SYCN | PGLYRP2 | CLCA1 | ASPHD1 | MS4A12 | GDPD3 | FAM57B | C16orf92 | CELA3B |
| SLC4A4 | AHCYL1 | CA2 | CA9 | CA4 | AHCYL2 | SLC26A6 | SLC9A3 | SLC9A1 | SLC12A2 | CFTR |
| CEACAM7 | NXPE4 | ZNF574 | TECTB | ANPEP | CYP2A7 | CYP2A13 | LRP6 | ICAM2 | CLCA4 | |
| ADH1C | CYP2E1 | ALDH2 | ADH1B | LRAT | AOX1 | CYP26A1 | RETSAT | BCO1 | CYP27C1 | PNPLA4 |
| DHRS9 | RBP1 | ALDH1A3 | ALDH1A1 | CYP26A1 | ALDH1A2 | LRAT | RBP2 | BCO1 | RETSAT | DGAT1 |
| PKIB | MPPED1 | SLAIN2 | AKAP14 | CALML5 | PKIG | TBATA | GREB1 | CDK10 | AKAP4 | NUDCD2 |
| CHP2 | SLC9A1 | PPP3CA | C9orf78 | SLC9A3 | PPP3CB | PPP3CC | FKBP3 | FKBP1C | FKBP1B | FKBP1A |
| CPM | BDKRB1 | YEATS4 | CPSF6 | ENPP3 | C12orf71 | KNG1 | SPRYD4 | JOSD2 | LYZL4 | BDKRB2 |
| HSD17B2 | HSD17B3 | HSD17B1 | SULT1E1 | CYP19A1 | SRD5A1 | HSD3B1 | HSD3B2 | CYP17A1 | SRD5A2 | DHRS11 |
| HPGD | HSD11B2 | SLCO2A1 | PTGES | RSPO4 | PTGS2 | PTGR1 | RGS12 | ZADH2 | PTGER4 | FZD6 |

Believing that cancer cells originate from adenomas, we can understand how the adenoma progresses to cancer if we identify the differences in gene expression between the two groups. **Table 5** shows the only genes that had the same gene expression in four tests of cancer cell comparisons with normal but they do not show any gene expression change in Test 5.

*Comparing of gene expression levels in different stages of CRC with normal*

The data obtained from microarray data analysis related to different stages of CRC cancer that were collected in GSE21510 [14] were examined. The results of this study that show the comparison of gene expression in normal cells with cancer cells in the first stage to the fourth stage for genes CLCA1 and SELENBP1, is shown in **Table 6**.

## Discussion

In the current study, we obtained gene expression signatures associated with the transition from adenoma to cancer in patients with CRC. Analysis of the GSE21510, GSE4107, GSE25071, GSE15781 and GSE8671 datasets identified 17 downregulated genes in CRC. On top of that, CLCA4 has the most reduction.

CLCA4 is known to be a tumor suppressor gene [19, 20], but in this study we found that this gene in adenoma also reduced its expression almost as much as cancer cells, indicating that it could not be used to detect cancer in early stage. However, as shown in **Table 1**, an 88-fold reduction in its expression in cancer cells could pave the way for cell cancer. Therefore, by studying which other molecules this gene affects, it can help us to understand which molecules can cause cancer in cells. For this purpose, PPI network was made for this molecule and other molecules in **Table 1**. As can be seen from **Figure 1**, the molecules in the PPI network of CLCA4 are somehow connected to the CLCA4 as co-expression connectivity type. As it is obvious in **Table 2**, some of the molecules in this network with this type of relationship show a similar trend. CLCA1 and MS4A12 have this kind of connection, and they, like CLCA4, show a greater reduction in expression than control, which suggests that these three molecules most likely have a strong influence on each other. Some of them show the opposite trend, for example, CWC25 spliceosome associated protein homolog (CWC25), which is connected to the CLCA4 and its six subsidiaries, has a declining trend, but most of its subsidiaries including CWC22, SNRNP200, PRPF19 and CDC5L have an increasing trend. From this inverse similarity it can be inferred that this decrease in CWC25 expression may be related to the increase in the expression of its subsets. Of course, since CWC25 does not show any difference in adenoma, it can be assumed that first its subsets underwent expression changes, then after passing through adenoma to cancer, we will see a decrease in CWC25 expression.

**Table 4.** Comparison of Gene expression in CRC with normal (Test 1 to Test 4) and adenoma with normal (Test 5)

| Comparison of gene expression of genes in PPI networks | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gene | Gene in PPI network | LogFC | | | | | | Connectivity Type |
| | | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average | |
| AQP8 | AQP9 | 0.1 | 0.3 | 3.4 | 2.3 | 1.3 | 1.5 | |
| | ARHGAP17 | -1.3 | -0.3 | -0.3 | -0.7 | -0.9 | -0.7 | |
| CA1 | SELENBP1 | -3.2 | -1.6 | -2.8 | -1.8 | | -2.4 | Co-Expression |
| GUCA2A | GUCY2C | -0.1 | -2.0 | -0.8 | -0.4 | | -0.9 | Co-Expression |
| | PDZD3 | -1.8 | -1.3 | -1.3 | -2.0 | -2.6 | -1.8 | Co-Expression |
| | SLC26A3 | -2.0 | -1.2 | -5.6 | -4.6 | -4.5 | -3.6 | Co-Expression |
| | CLCA1 | -4.0 | -2.4 | -5.1 | -2.6 | | -3.5 | Co-Expression |
| CA2 | SLC4A4 | -4.0 | -1.7 | -4.7 | -3.8 | -3.3 | -3.5 | |
| | CDH1 | -1.0 | -1.6 | -0.5 | -0.5 | -0.3 | -0.8 | |
| CD177 | PROCR | 2.7 | 0.3 | 1.5 | 0.7 | | 1.3 | |
| ZG16 | CLCA1 | -4.0 | -2.4 | -5.1 | -2.6 | | -3.5 | Co-Expression |
| | MS4A12 | -7.1 | -1.5 | -6.4 | -4.6 | -5.9 | -5.1 | Co-Expression |
| | GDPD3 | -3.1 | -1.4 | -1.6 | -1.4 | -1.4 | -1.8 | Co-Expression |
| SLC4A4 | CA4 | -6.9 | -1.5 | -4.6 | -4.6 | -5.5 | -4.6 | Co-Expression |
| | AHCYL2 | -2.7 | -1.8 | -2.7 | -2.0 | -1.8 | -2.2 | Co-Expression |
| | SLC9A3 | -0.6 | -2.0 | -0.6 | 0.0 | -1.8 | -1.0 | |
| CEACAM7 | NXPE4 | -3.1 | -1.5 | -4.8 | -2.8 | -1.6 | -2.8 | Co-Expression |
| | CLCA4 | -7.9 | -1.7 | -7.1 | -4.3 | -5.3 | -5.2 | Co-Expression |
| ADH1C | RETSAT | -2.8 | -1.8 | -1.4 | -1.2 | -1.2 | -1.7 | |
| DHRS9 | ALDH1A1 | -1.6 | -0.1 | -1.3 | -0.3 | | -0.8 | Co-Expression |
| | RETSAT | -2.8 | -1.8 | -1.4 | -1.2 | -1.2 | -1.7 | |
| | DGAT1 | -1.1 | -0.8 | -0.5 | -0.3 | -0.5 | -0.6 | |
| CHP2 | SLC9A3 | -0.04 | -2.0 | -0.6 | -0.05 | -1.8 | -0.9 | |
| CPM | YEATS4 | 2.7 | 0.2 | 0.8 | 1.2 | 1.0 | 1.2 | |
| | BDKRB2 | -1.8 | -0.2 | -0.8 | -0.5 | -0.9 | -0.8 | |
| HSD17B2 | DHRS11 | -3.6 | -1.5 | -2.0 | -2.1 | -2.4 | -2.3 | |
| HPGD | HSD11B2 | -3.3 | -1.7 | -1.9 | -1.8 | -1.4 | -2.0 | Co-Expression |
| | PTGS2 | 1.6 | 2.7 | 1.3 | 2.5 | | 2.0 | |
| | PTGER4 | -1.7 | -1.2 | -2.0 | -0.9 | -0.5 | -1.2 | |
| | FZD6 | 1.8 | 0.2 | 1.1 | 1.1 | 0.8 | 1.0 | |

Also, by continuing this method of studying on the other genes in **Table 1**, creating a PPI network and studying the expression of each of them in all 5 Tests, we identified 30 deregulated genes (**Table 4**). Notable points in **Table 4** include the co-expression type of connectivity between CLCA1 with guanylate cyclase activator 2A (GUCA2A) and zymogen granule protein 16 (ZG16), and CLCA4 with carcinoembryonic antigen related cell adhesion molecule 7 (CEACAM7); so, these proteins can be considered very important for this disease as well. The importance of GUCA2A in preventing clone cancer [21], downregulation of ZG16 [22] and CEACAM7 [23] in CRC have been proven.

Interestingly, by comparing the expression genes in normal, cancer and adenoma cells in this study, 10 of them (CLCA1, selenium binding protein 1 (SELENBP1), CWC25, acyl-CoA thioesterase 11 (ACOT11), guanylate cyclase 2C (GUCY2C), aldehyde dehydrogenase 1 family member A1 (ALDH1A1), prostaglandin-endoperoxide synthase 2 (PTGS2), protein C receptor (PROCR), molybdenum cofactor synthesis 3 (MOCS3) and NFS1 cysteine desulfurase (NFS1)) have been identified as important genes in

**Table 5.** Genes with decreased or increased gene expression compared CRC with normal, and no change in gene expression in adenoma cells with normal (Test 5)

| Comparison of gene expression | | | | | | | |
|---|---|---|---|---|---|---|---|
| Gene Symbol | LogFC | | | | | | Comment |
| | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average | |
| CLCA1 | -4.0 | -2.4 | -5.1 | -2.6 | | -3.5 | Suppressor genes |
| SELENBP1 | -3.2 | -1.6 | -2.8 | -1.8 | | -2.4 | |
| CWC25 | -0.2 | 0.0 | -3.2 | -0.1 | | -0.9 | |
| ACOT11 | -1.5 | -1.0 | -0.5 | -0.5 | | -0.9 | |
| GUCY2C | -0.1 | -2.0 | -0.8 | -0.4 | | -0.9 | |
| ALDH1A1 | -1.6 | -0.1 | -1.3 | -0.3 | | -0.8 | |
| PTGS2 | 1.6 | 2.7 | 1.3 | 2.5 | | 2.0 | oncogene |
| PROCR | 2.7 | 0.3 | 1.5 | 0.7 | | 1.3 | |
| MOCS3 | 0.1 | 0.6 | 1.5 | 0.8 | | 0.8 | |
| NFS1 | 0.8 | 0.1 | 0.3 | 0.4 | | 0.4 | |

**Table 6.** The comparison of gene expression of highly downregulated genes in **Table 5** (CLCA1 and SELENBP1) in different stages of CRC with normal

| Gene expression of CLCA1 and SELENBP1 in different stages of CRC | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Genes | Log2FC | | | | Fold Change | | | |
| | Stage 1 | Stage 2 | Stage 3 | Stage 4 | Stage 1 | Stage 2 | Stage 3 | Stage 4 |
| CLCA1 | -3.2 | -4.2 | -4.5 | -3.4 | -9.3 | -18.1 | -22.6 | -10.8 |
| SELENBP1 | -2.0 | -3.6 | -3.4 | -2.8 | -4.0 | -11.8 | -10.7 | -7.1 |

the transition from adenoma to cancer; due to they do not show any change in gene expression in adenoma but in cancer, suggesting to be considered as biomarker for early detection and treatment in CRC. CLCA1 was studied and considered as a suppressor gene [24, 25] in CRC but here we reported a change in its expression from adenoma to cancer, we saw an average of 15.7 fold decrease in expression of this gene in CRC in all 4 tests while no change between normal and adenoma. Downregulation of SELENBP1 [26], GUCY2C [27], ALDH1A1 [28] and upregulation of PTGS2 [29], PROCR [30], NFS1 [31] in comparison CRC with normal cells have been revealed. Among the 10 mutable genes in expression in the transition from adenoma to cancer, CLCA1 and SELENBP1 with a decrease in expression of 15.7 and 5.3 times, respectively, and PTGS2 with a 4-fold increase, each in the progression of this disease is very important. According to the co-expression type of PPI connectivity between CLCA1 with CLCA4, the head of the PPI, with a decrease means of 88 times, and SELENBP1 with CA1, the head of the PPI network, which shows a 61-fold decrease in expression, it creates deep beliefs that: 1. Downregulation of CLCA1 and SELENBP1 from adenoma to cancer probably originate from the downregulation of CLCA4 and CA1 and 2. CLCA1 and SELENBP1 could be more important for early detection of CRC. CLCA1 mRNA expression level has been shown an inversely correlation with CRC stage and metastasis and in vitro and in vivo experiments have proved that upregulation of CLCA1 suppressed CRC metastasis and growth and its downregulation led to the opposite results [25]. Furthermore, the researchers discovered that a decrease in SELENBP1 expression in CRC lowers the overall survival rate, and that suppressing this gene hastens the disease's progression [32]. Totally, among novel six and four-gene signatures as suppressors and oncogene genes respectively as independent predictors that we conceived for early detection of CRC, although further validations are required for using these signature genes in the clinic, CLCA1 and SELENBP1 are best in this case.

## Disclosure of conflict of interest

None.

**Address correspondence to:** Bahram Goliaei, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. Tel: +98-21-6649-8672, +98-912-123-9435; E-mail: goliaei@ut.ac.ir; faalashti@ut.ac.ir

## References

[1] Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, Cercek A, Smith RA and Jemal A. Colorectal cancer statistics, 2020. CA Cancer J Clin 2020; 70: 145-164.

[2] Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A and Jemal A. Colorectal cancer statistics, 2017. CA Cancer J Clin 2017; 67: 177-193.

[3] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.

[4] Nilsson R, Bjorkegren J and Tegner J. On reliable discovery of molecular signatures. BMC Bioinformatics 2009; 10: 38.

[5] Zong Z, Li H, Yi C, Ying H, Zhu Z and Wang H. Genome-wide profiling of prognostic alternative splicing signature in colorectal cancer. Front Oncol 2018; 8: 537.

[6] Bertucci F, Salas S, Eysteries S, Nasser V, Finetti P, Ginestier C, Charafe-Jauffret E, Loriod B, Bachelart L, Montfort J, Victorero G, Viret F, Ollendorff V, Fert V, Giovaninni M, Delpero JR, Nguyen C, Viens P, Monges G, Birnbaum D and Houlgatte R. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene 2004; 23: 1377-1391.

[7] Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ and Tollenaar R. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol 2011; 29: 17-24.

[8] Data Mining Procedures Using GEO (Gene Expression Omnibus). China Biotechnology 2007; 27: 96-103.

[9] Hu G and Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One 2009; 4: e6536.

[10] Titeca K, Lemmens I, Tavernier J and Eyckerman S. Discovering cellular protein-protein interactions: technological strategies and opportunities. Mass Spectrom Rev 2019; 38: 79-111.

[11] Herce HD, Deng W, Helma J, Leonhardt H and Cardoso MC. Visualization and targeted disruption of protein interactions in living cells. Nat Commun 2013; 4: 2660.

[12] Barh D, Yiannakopoulou EC, Salawu EO, Bhattacharjee A, Chowbina S, Nalluri JJ, Ghosh P and Azevedo V. Chapter 22-In silico disease model: from simple networks to complex diseases. In: Verma AS, Singh A, editors. Animal Biotechnology (Second Edition); 2020. pp. 441-460.

[13] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ and Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019; 47: D607-D613.

[14] Tsukamoto S, Ishikawa T, Iida S, Ishiguro M, Mogushi K, Mizushima H, Uetake H, Tanaka H and Sugihara K. Clinical significance of osteoprotegerin expression in human colorectal cancer. Clin Cancer Res 2011; 17: 2444-2450.

[15] Hong Y, Ho KS, Eu KW and Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin Cancer Res 2007; 13: 1107-1114.

[16] Agesen TH, Berg M, Clancy T, Thiis-Evensen E, Cekaite L, Lind GE, Nesland JM, Bakka A, Mala T, Hauss HJ, Fetveit T, Vatn MH, Hovig E, Nesbakken A, Lothe RA and Skotheim RI. CLC and IFNAR1 are differentially expressed and a global immunity score is distinct between early- and late-onset colorectal cancer. Genes Immun 2011; 12: 653-662.

[17] Snipstad K, Fenton CG, Kjaeve J, Cui G, Andersen E and Paulssen RH. New specific molecular targets for radio-chemotherapy of rectal cancer. Mol Oncol 2010; 4: 52-64.

[18] Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E, Maake C, Rehrauer H, Laczko E, Kurowski MA, Bujnicki JM, Menigatti M, Luz J, Ranalli TV, Gomes V, Pastorelli A, Faggiani R, Anti M, Jiricny J, Clevers H and Marra G. Transcriptome profile of human colorectal adenomas. Mol Cancer Res 2007; 5: 1263-1275.

[19] Wei L, Chen W, Zhao J, Fang Y and Lin J. Downregulation of CLCA4 expression is associated with the development and progression of colorectal cancer. Oncol Lett 2020; 20: 631-638.

[20] Liu Z, Chen M, Xie LK, Liu T, Zou ZW, Li Y, Chen P, Peng X, Ma C, Zhang WJ and Li PD. CLCA4 inhibits cell proliferation and invasion of hepatocellular carcinoma by suppressing epithelial-mesenchymal transition via PI3K/AKT signaling. Aging (Albany NY) 2018; 10: 2570-2584.

[21] Pattison AM, Merlino DJ, Blomain ES and Waldman SA. Guanylyl cyclase C signaling axis and

colon cancer prevention. World J Gastroenterol 2016; 22: 8070-8077.
[22] Meng H, Li W, Boardman LA and Wang L. Loss of ZG16 is associated with molecular and clinicopathological phenotypes of colorectal cancer. BMC Cancer 2018; 18: 433.
[23] Bian Q, Chen J, Qiu W, Peng C, Song M, Sun X, Liu Y, Ding F, Chen J and Zhang L. Four targeted genes for predicting the prognosis of colorectal cancer: a bioinformatics analysis case. Oncol Lett 2019; 18: 5043-5054.
[24] Bustin SA, Li SR and Dorudi S. Expression of the Ca2+-activated chloride channel genes CLCA1 and CLCA2 is downregulated in human colorectal cancer. DNA Cell Biol 2001; 20: 331-338.
[25] Li X, Hu W, Zhou J, Huang Y, Peng J, Yuan Y, Yu J and Zheng S. CLCA1 suppresses colorectal cancer aggressiveness via inhibition of the Wnt/beta-catenin signaling pathway. Cell Commun Signal 2017; 15: 38.
[26] Wang N, Chen Y, Yang X and Jiang Y. Selenium-binding protein 1 is associated with the degree of colorectal cancer differentiation and is regulated by histone modification. Oncol Rep 2014; 31: 2506-2514.
[27] Bashir B, Merlino D, Rappaport J, Gnass ED, Palazzo J, Fing Y, Fearon ER, Snook A and Waldman SA. Guanylate cyclase C (GUCY2C) as a preventative and therapeutic target in colorectal cancers (CRCs) arising through divergent genomic pathways. J Clin Oncol 2019; 37 Suppl: 595-595.
[28] Matsumoto A, Arcaroli J, Chen Y, Gasparetto M, Neumeister V, Thompson DC, Singh S, Smith C, Messersmith W and Vasiliou V. Aldehyde dehydrogenase 1B1: a novel immunohistological marker for colorectal cancer. Br J Cancer 2017; 117: 1537-1543.
[29] Zahedi T, Hosseinzadeh Colagar A and Mahmoodzadeh H. PTGS2 over-expression: a colorectal carcinoma initiator not an invasive factor. Rep Biochem Mol Biol 2021; 9: 442-451.
[30] Lal N, Willcox CR, Beggs A, Taniere P, Shikotra A, Bradding P, Adams R, Fisher D, Middleton G, Tselepis C and Willcox BE. Endothelial protein C receptor is overexpressed in colorectal cancer as a result of amplification and hypomethylation of chromosome 20q. J Pathol Clin Res 2017; 3: 155-170.
[31] Wang FT, Li XP, Pan MS, Hassan M, Sun W and Fan YZ. Identification of the prognostic value of elevated ANGPTL4 expression in gallbladder cancer-associated fibroblasts. Cancer Med 2021; 10: 6035-6047.
[32] Kim H, Kang HJ, You KT, Kim SH, Lee KY, Kim TI, Kim C, Song SY, Kim HJ, Lee C and Kim H. Suppression of human selenium-binding protein 1 is a late event in colorectal carcinogenesis and is associated with poor survival. Proteomics 2006; 6: 3466-3476.