Project 1 Group wheeler Roles:

Data Curator: Ariel Xue

Programmer: Jessica Fetterman

Analyst: Reina Chau

Biologist: Vishala Mishra

Introduction

Colon cancer is the third most common cancer in the world. Although advanced technologies have improved the screening, diagnosis, and treatment of colon cancer, colon cancer continues to be a major cause of mortality. There are five pathological stages of colon cancer, and the prognosis and treatment process for each patient depends greatly on the identification of the stage. Patients in stages II and III have a higher rate of recurrence. The goal of this study is to improve colon cancer stratification and predict prognosis based on gene expression profiles using a microarray technology. Gene expression data from microarrays are suitable for this study because the microarrays can efficiently show which genes are differentially expressed between colon cancer cells and normal cells by comparing the abundance of mRNAs at specific genomic loci. The underlying associations between cancer and normal cells are evaluated with a Chi-square test. To ensure the accuracy of classification, samples are divided into two clusters using unsupervised hierarchical clustering, then a Welch t-test was applied to identify genes differentially expressed between the two clusters.

Methods

Study Samples

Ethics committees approved tumor tissue samples collection on 750 stage I to IV colon cancer patients was performed on medical records from several French medical institutions. No patient with preoperative chemotherapy or radiation therapy was included in the study. American Joint Committee on Cancer tumor node metastasis staging system was used as a protocol to stage each patient. Genome samples from human primary colorectal Adenocarcinoma are determined on Affymetrix U133 Plus 2.0 chips for gene expression analysis data. Of the 750 samples, 566 samples were considered to be qualified for gene expression analysis. Array-based comparative genomic hybridization could be performed on 464 out of 750 primary samples with bacterial artificial chromosome (BAC) microarrays. 443 samples are put into the discovery set and 123 samples into the validation set.

An additional 906 samples from public datasets are included in the validation set and they are also obtained from similar chip platforms with tumor location and DNA alteration. RNA qualities are controlled by stringent criteria - 28s/18s ratio above 1.8 for microarray to rule out degraded data. All the hybridization and amplification processes followed the manufacturer's protocol to ensure the high quality of data. Thus, only the high-quality part of the samples (566 out of 750) went through gene expression analysis. No contamination was specified on the collection of data.The 750 samples come from the French national Cares d'Identite' des Tumeurs (CIT) program between 1987 to 2007.[1]

Creation of the Dataset

All of the programming was run using R Studio on the SCC server. I created an R script (/projectnb/bf528/users/wheeler/project_1/analysis/programmer/project_1_programmer.R) that I ran in the console in R Studio (version 3.5.1), saving the csv files to the group's analyst folder (/projectnb/bf528/users/wheeler/project_1/analysis/analyst/). I started by installing and loading Bioconductor and the BiocManager, affy, sva, affyPLM, AnnotationDbi, hgu133plus2.db, ggplot2, and tidyverse packages in RStudio on the SCC Server. It took some time to install Bioconductor and the packages initially but then loading them took very little time.

Next, I read the CEL files using the ReadAffy() function in the affy package with the celfile.path specifying the directory containing all 134 CEL files. The ReadAffy function reads the CEL files into an Affybatch that holds all the information on the phenoData (an annotated data frame), with the row names being the names of the CEL files and a column with the samples. I then used the rma() function in the affy package to normalize all of the CEL files. The rma() function converts the data into an ExpressionSet, applying quantile normalization and a background correction. The rma() function assumes the background noise is normally distributed across the microarray and subtracts it from the sample data, much like subtracting a negative control in an experiment where the measure is taken from the assay reagents in the absence of a sample. The output consists of an expression measure in log base 2 scale, which I saved under the variable name RMA_Data containing 134 samples (columns) with 54,675 features (these are the gene expression values and are in the rows).I wrote the RMA_Data to a csv file as a table with the following path-

/projectnb/bf528/users/wheeler/project_1/analysis/analyst/RMA_Data.csv

Evaluation of Sample Quality

I calculated the Relative Log Expression (**RLE**) and Normalized Un-scaled Standard Error (**NUSE**) scores in order to assess the quality of the data. The fitPLM function converts an AffyBatch object (e.g., ReadAffyData) to a PLM dataset, while also performing a quantile normalization and RMA background correction. Because this step was re-applying the RMA background correction, I used the Affybatch data created from the ReadAffy() function, rather than the ExpressionSet, in order to not repeat the rma() background correction. Of note, this step took some time to run (e.g., ~5 minutes or so).
Next, I calculated the median RLE using the RLE() function with the argument type=”stats”, which computes the medians and the interquartile ranges for each microarray sample. I indexed the medians (row 1) and saved it as the variable median. Turns out, ggplot only accepts variables from a data frame so I converted variable to a data frame with the data.frame function. I used ggplot (loaded in the tidyverse package) to create a histogram of the median RLE’s, a data summarization method.

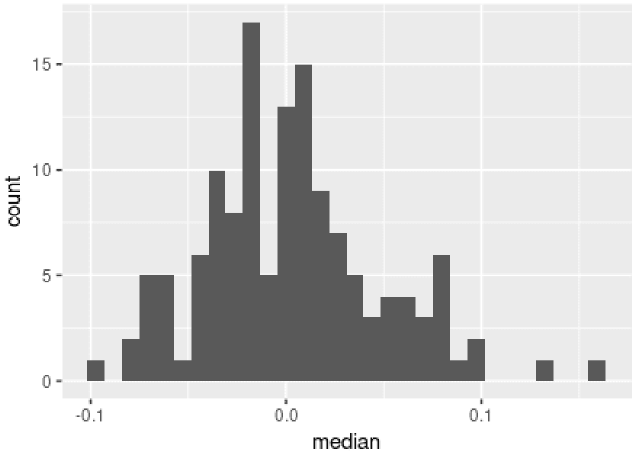

Figure 1. Median RLEs for the validation samples

RLE is a quality assessment tool where most of the RLE values for the genes on an array, if high quality, should ideally be centered around 0, which they largely are for this dataset (**Figure 1**). Several of the samples fall to the left and right of the somewhat normal distribution of median log expression and may be outliers.

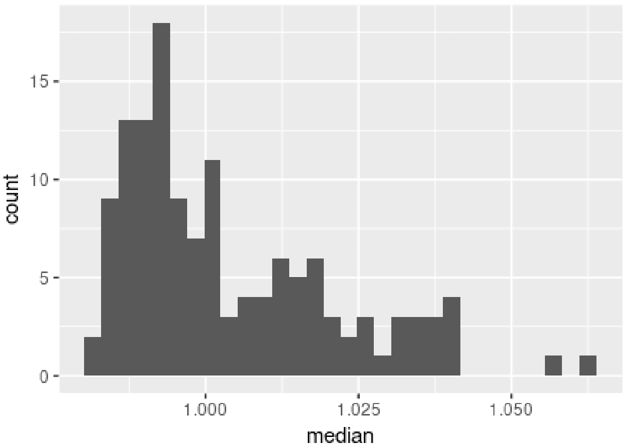

Figure 2. Median NUSE for the validation samples

Similarly, I calculated the median NUSE using the NUSE() function with the argument type=”stats”, which computes the medians and interquartile ranges for each microarray sample. Again, I indexed the medians (row 1) and saved it as the variable median. I converted the row of median values to a data frame and then used ggplot to create a histogram of the median NUSE for each sample.

NUSE is another metric of quality for microarray data. The standard error estimates should fall around 1, if the samples are of high quality. In this case, there are peaks (high counts) around 1 but there’s more of a spread to the right (**Figure 2**). It is possible the skew in the distribution of the majority of the samples could be reflective of the presence/absence of disease or the severity of the disease (e.g., in this study, the stage of cancer). A few of the samples appear to be outliers because they fall really far to the right of the distribution of median NUSE, which means they have a really high variation that could be due to technical issues in sample preparation or running.

I used the ComBat() function in the sva package to correct for batch effects while maintaining the features of interest. The batch correction is important for “removing” the variance related to confounding variables such as the time the microarray was run (e.g., which batch of samples), RNA extraction method, and other confounders related to the processing of the samples. The features of interest include the outcomes of interest such as stage of cancer and relapse score classification. The ComBat() function requires several arguments including the dataset (rma normalized data), the batch effect variable (normalizationcombatbatch), and the features of interest variable (normalizationcombatmod). The features of interest variable had to be converted to a matrix as the ComBat() function only accepts a model matrix for this argument. Finally, I wrote the batch corrected expression values to a csv file:

/projectnb/bf528/users/wheeler/project_1/analysis/analyst/batch_corr_data_final.csv

Principal Component Analysis

I performed a Principal Component Analysis (**PCA**) on the normalized data from step 5 using the prcomp() function. First, I had to scale and center the columns by gene, which required that I transpose the matrix with the t() function. The scaling takes into account that the measures may have different units and transforms the data to the same scale so that we’re comparing “apples to apples” so-to-speak. Next, I applied the prcomp() function to perform the PCA, setting the scale and center arguments to FALSE since I already scaled and centered the genes. Had I performed the scaling and centering as part of the prcomp() function, the program would have scaled and centered the samples, not the genes. I used the summary() function to obtain the amount of variance in the expression data attributed to the first two principal components (**PCs**). PC1 explains 94.79% of the variation in gene expression and PC2 explains 1.133% of the variation in the gene expression. So most of the variation is explained in the first PC.

I plotted PC1 and PC2 using the ggbiplot() function from the tidyverse and ggplot2 packages. The ggbiplot() function takes the prcomp function output as an argument and creates a scatterplot from PC1 and PC2 as the default, although there’s an option to plot different PC’s. I had a hard time getting this function to work.

A PCA is a dimensionality reduction tool that identifies the features that collectively contribute the most to the variation in the gene expression. In this study, I found that the PC1 explains 95% of the variance in the microarray samples and PC2 explains % of the variance in gene expression of the microarray samples. Collectively, PC1 and PC2 explain 19.9% of the total variance.

Overall, the greatest challenge I had while working on the project was all of the different data types, particularly the multi-layered ones such as the GeneExpression set. All of the different types of matrices and what functions could be passed on each really created some headaches. I've learned through the process that I just need to check the data type with class() and the data format for arguments when putting together my code. Specifically, I have really struggled with visualizing the multi-dimensional data types, which really made it hard for me to figure out how to index and extract specific pieces of information. It took me a long time to figure out just what I was looking at with the multi-layer datasets and how to figure out what all was inside the matrix.

Noise Filtering and Dimensionality Reduction

The "large p, small n" is a well-known problem that arises in microarray studies, where p is the number of features (e.g. genes) is much larger than the number of samples (n, e.g. patients). Therefore, many methods can be used to remove this noise. In this study, we adopted three filtering metrics to reduce the number of features in the normalized gene expression set. First, the genes were filtered by the criteria that at least 20% of the gene expression-values must be > log2(15). Second, a chi-squared test was performed to assess whether there is a significant difference between a gene's variance and the overall median variance of all probe sets. Genes that do not pass the threshold of $p < 0.01$ will be removed from the gene expression set. Lastly, a coefficient of variation (CV) was calculated to measure the level of dispersion around the mean. Only genes with CV values > 0.186 will be retained.

Hierarchical Clustering and subtype discovery

In this study, we are only interested in the tumor classification of C3 and C4 subtypes. In order to determine the subtype membership for each patient without relying on its class labels, an unsupervised hierarchical clustering was performed using Ward linkage and Euclidean distance. The clustering was done on the patient levels in which the patient samples were divided into two clusters that roughly resemble the subtypes of C3 and C4. A Welch t-test was computed to identify genes that are differentially expressed between the two clusters, and the Benjamin Hochberg (FDR) method was used to correct for multiple testings. Likewise, to classify genes that are associated with each subtype, an unsupervised hierarchical clustering was performed using Ward linkage and 1 - Pearson correlation coefficient distance. Both the subtype and gene memberships were visualized on a heatmap where the subtype clustering is represented by the column dendrogram and the gene clustering is represented by the row dendrogram.

Enrichment Analysis

KEGG, GO, and cancer hallmark gene sets were obtained from MSigDB and used to evaluate the biological significance of the different gene expression profiles for each tumor subtype. There are 186 gene sets in the KEGG pathways, 10,271 gene sets in Gene Ontology, and 50 gene sets in cancer hallmark. A mapping from a probeset ID to a gene symbol was done using the bioconductor package **hgu133plus2.db**. Out of 47,334 probeset IDs, there are 8,914 probe sets that do not map to any gene symbols which is quite concerning. I tried to remove leading and trailing whitespace in the probeset IDs but there is no effect. A further investigation must be conducted to understand why NAs were produced during the mapping. Nevertheless, with the rest of the probe sets that were successfully mapped to a gene symbol, a Fisher's Exact test was performed to associate the gene sets to cancer signatures, the biological process, cellular component, molecular function, and KEGG pathways for each tumor subtype. Benjamini-Hochberg (FDR) method was used to correct for multiple testings.

Results

The three filtering metrics were applied to reduce the number of features in the normalized data. The first filter reduced the total genes of 54,675 to 39,661 genes. The second filter using a chi-square test lowered the total number of genes to 29,645 genes. Lastly, the third filter significantly brought the total number of genes to 1,531 genes.

With a gene expression set of 1,531 genes, an unsupervised classification was used to explore its subtype and gene memberships. For the subtype classification, the hierarchical clustering was able to classify the 134 patients into two clusters, with 55 samples for cluster 1 and 77 samples for cluster 2. The heatmap in **Figure 3** shows the column dendrogram of subtype memberships based on the gene expression of 1531 probe sets. A Welch t-test was computed to compare the differentially expressed genes between the two clusters. Out of 1531 genes, 1236 genes were differentially expressed between the two clusters using Benjamin Hochberg (FDR) method with adjusted p < 0.05. Within 1,236 differentially expressed genes, we identified 909 genes that are highly expressed to the cluster 1, and 327 genes that are highly expressed to cluster 2. The row dendrogram in **Figure 3** shows a list of genes that are associated with each subtype using Ward linkage and 1 - Pearson correlation coefficient distance. To further validate our findings, we looked at the top 1000 up- and down-regulated genes with its positive and negative log2 fold change. The top 10 of these up- and down-regulated genes are shown in **Table 1**.

KEGG, GO, and cancer hallmark gene sets are used to evaluate biological significance that profiles each tumor subtype. Based on the Fisher's Exact test with Benjamin Hochberg (FDR) correction (adjusted p < 0.05), there are 7 gene sets enriched in KEGG, 120 gene sets enriched for GO, and 24 gene sets enriched for hallmark. The top 3 enriched gene sets for each geneset type are shown in **Table 2**.

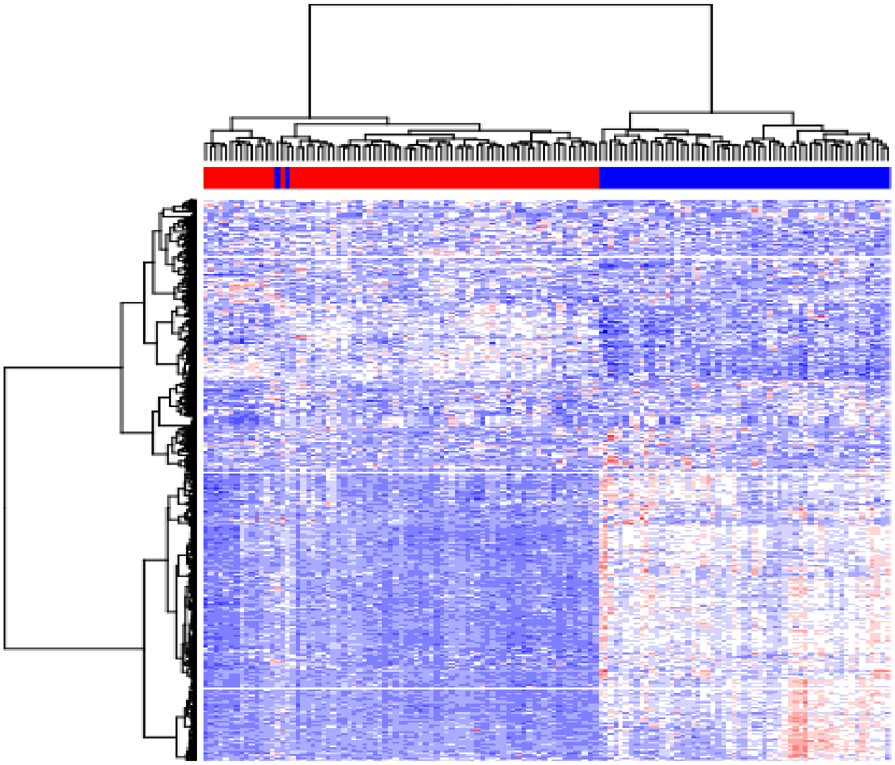

Figure 3: Heatmap of 1531 probe sets ordered by tumor subtypes. The red box represents the C3 subtype (n=75), and the blue box represents the C4 subtype (n=59). The row dendrogram provides information about the gene associations using Ward linkage and 1 - Pearson correlation coefficient distance. The column dendrogram provides information about the subtype memberships using Ward linkage and Euclidean distance.

Gene Symbol	Direction	Log2 Fold Change	T	P	Adjusted P
FNDC1	Up	0.973	21.741	< 0.0001	< 0.0001
MIR100HG	Up	0.953	17.908	< 0.0001	< 0.0001
PLN	Up	0.922	17.835	< 0.0001	< 0.0001
EPYC	Up	0.890	6.978	< 0.0001	< 0.0001
SFRP4	Up	0.876	17.883	< 0.0001	< 0.0001
SFRP2	Up	0.875	21.980	< 0.0001	< 0.0001
LINC01279	Up	0.802	21.208	< 0.0001	< 0.0001
HTR2B	Up	0.786	10.803	< 0.0001	< 0.0001
NOX4	Up	0.780	16.099	< 0.0001	< 0.0001
ZFPM2	Up	0.768	18.566	< 0.0001	< 0.0001
HEPACAM2	Down	-0.930	-12.586	< 0.0001	< 0.0001
CLCA1	Down	-0.905	-13.274	< 0.0001	< 0.0001
ITLN1	Down	-0.762	-11.098	< 0.0001	< 0.0001
FCGBP	Down	-0.744	-16.219	< 0.0001	< 0.0001
SI	Down	-0.706	-7.106	< 0.0001	< 0.0001
SPINK4	Down	-0.679	-12.212	< 0.0001	< 0.0001
L1TD1	Down	-0.660	-10.477	< 0.0001	< 0.0001
CLCA4	Down	-0.627	-7.725	< 0.0001	< 0.0001
RETNLB	Down	-0.622	-9.644	< 0.0001	< 0.0001
CEACAM7	Down	-0.618	-7.534	< 0.0001	< 0.0001

Table 1 - Top 10 up and down-regulated genes were filtered by the positive and negative log2 fold change between the two clusters and the chi-squared test is used to identify differentially expressed genes with Benjamini-Hochberg correction for multiple testing (adjusted p < 0.01)

Collection	Gene Sets	Statistics	P
Hallmark	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	3.339	1E-13
	HALLMARK_FATTY_ACID_METABOLISM	2.388	2E-06
	HALLMARK_ADIPOGENESIS	2.130	3E-06
KEGG	KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM	8.946	3E-06
	KEGG_ECM_RECEPTOR_INTERACTION	2.603	9E-05
	KEGG_PENTOSE_AND_GLUCURONATE_INTERCONVERSIONS	8.798	2E-04
GO	GO_CIRCULATORY_SYSTEM_DEVELOPMENT	1.613	1E-13
	GO_VASCULATURE_DEVELOPMENT	1.774	7E-12
	GO_TUBE_MORPHOGENESIS	1.642	2E-11

Table 2 - Top 3 enriched gene sets for KEGG, GO, and Hallmark collections.

Discussion

The purpose of the study was to identify gene expression patterns in primary tumor samples that improve the risk stratification for colon cancer recurrence. Our filtered gene probe dataset consisted of 1,531 probe sets, which is similar to the total number of probe sets in the publication. Using hierarchical clustering, we identified two clusters of gene expression patterns, C3 and C4 subtypes, in the validation cohort used for the publication.

A chi square test revealed that among the top 10 genes that delineate the two clusters, several up-regulated genes (*SFRP4*, *SFRP2*) play important roles in regulating Wnt signaling, which is consistent with the C4 cluster identified by Marisa L. *et al* that are characterized by higher expression of Wnt signaling genes, and was associated with a worse prognosis.[1] Additionally, we identified NOX4 as being up-regulated in one of the clusters, which has relevance to a more pro-inflammatory phenotype since this gene encodes an enzyme, NADPH oxidase 4, that generates reactive oxygen species in order to destroy pathogens phagocytosed by immune cells like macrophages. The presence of higher *NOX4* expression suggests that the tumor samples belonging to this cluster have a high immune cell infiltration. Several of the top 10 up-regulated genes delineating the two clusters we identified play roles in the extracellular matrix (*FNDC1*, *EPYC*) and one long-non-coding RNA (*MIR100HG*) plays a role in regulating cellular proliferation, which makes sense as tumors are essentially a cluster of cells that have lost the ability to regulating their proliferation.

Among the top 10 down-regulated genes delineating our two clusters, we identified several genes that are specifically expressed in the intestine (*SI*, *SPINK4*, *CLCA1, CLCA4, RETNLB*), which makes sense since these are colon cancer samples. Yet, this was surprising to find these genes relevant to intestinal function were down-regulated, which may reflect the transformation of normal intestinal cells to cancer cells. One gene, *CEACAM7*, has previously been shown to be down-regulated in colon cancer and a predictive marker of rectal cancer recurrence, which is consistent with our finding that one of the clusters of tumor samples has down-regulated *CEACAM7*.[2] Lastly, *HEPACAM2*, a gene that plays a role in regulating mitosis, was down-regulated in one of the tumor sample clusters we identified, which makes sense in that mitosis contains several breaks on cellular proliferation that if lost, can result in run-away division of the cell, resulting in a tumor.

In our pathways analysis, we identified pathways involved in epithelial-mesenchymal transition, fatty acid metabolism, extracellular matrix receptor interaction, pentose and glucoronate interconversions, and several pathways involved in creation of new blood vessels- all of which are highly relevant to the development of cancer. For example, colon cancer is often derived from the epithelial layer of the colon proliferating to excess to result in tumor formation and part of this transformation process involves the loss of interactions of the cells with the basement membrane, allowing the cells to spread.[3] Additionally, angiogenesis is important for providing sufficient blood supply to a growing tumor so it is unsurprising that genes in pathways involved in angiogenesis are enriched in the tumor samples. The publication by Marisa L *et al.* also found many of these sample pathways in their analysis including the angiogenesis pathways and metabolic pathways.

Our analyses utilized the samples in the validation dataset, consisting of 134 tumor samples; however the manuscript reported 123 tumor samples.[1] It is possible that the possible outliers noted in the NUSE and RLE evaluation of the quality of the microarray data should have been filtered out of our data set and may contribute to the larger number of tumor samples used in our study compared to the publication. While the paper presented data on 443 samples in their discovery set and identified 6 different clusters based upon gene expression patterns, we identified 2 clusters using the 134 samples used in the paper's validation dataset, which is likely just due to differences in the sample sizes with a greater sample size providing more power to detect smaller differences in gene expression variation related to the clinical phenotype.

There are some challenges that we encountered while analyzing the microarray data. For example, many of the probeset IDs are mapped to NAs while we were retrieving its gene symbols. Therefore, those probeset IDs were removed from further analysis. The removal of those probe sets could have an effect on our downstream analysis of gene set enrichment. As a result, that could be one of the reasons that we could see a different result between our findings and the author's. Nevertheless, a further investigation needed to be conducted to understand why NAs were produced during the mapping process, and we are still looking for a solution to resolve this issue.

Conclusion

In conclusion, by reproducing some of the results from the original paper by Marisa *et al.*, we found a list of genes that are differentially expressed between the two tumor subtypes, C3 or C4.[1] We used the differentially expressed genes to identify KEGG pathways, cancer signatures, and biological processes that are associated with the colon disease. Although some of the analysis we did are not the same as mentioned in the paper (Consensus clustering vs. hierarchical clustering), it is reassuring that most of our results are similar with the author's findings. The association of the clusters with clinical data suggests that gene expression profiling of tumors in colon cancer may provide information on the risk of recurrence, which would warrant more frequent screening in those patients with gene expression patterns indicative of greater risk of recurrence.

References

1. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Flejou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P and Boige V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*. 2013;10:e1001453.
2. Messick CA, Sanchez J, Dejulius KL, Hammel J, Ishwaran H and Kalady MF. CEACAM-7: a predictive marker for rectal cancer recurrence. *Surgery*. 2010;147:713-9.
3. Kalluri R and Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest*. 2009;119:1420-8.