SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

May 2014 Vol.57 No.5: 526–531
doi: 10.1007/s11427-014-4650-1

Screening and regulatory network analysis of survival-related genes of patients with colorectal cancer

QI Lu[1] & DING YanQing[1,2*]

[1]*Department of Pathology, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China;*
[2]*Department of Pathology, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China*

Received September 21, 2013; accepted November 17, 2013; published online April 9, 2014

The purpose of this study was to screen key survival-related genes from patients with colorectal cancer and explore signal transduction network of the involved genes. In a previous study, survival-related genes of patients with colorectal cancer were selected by colorectal cancer-related expression data GSE17538 using the Significance Analysis of Microarrays (SAM3.01) software, and 235 genes related to the survival of patients with colorectal cancer were obtained. Therefore, the following screening and analysis were conducted on these 235 genes in this study. First, the enrichment analysis of transcription factor binding sites was conducted on the 235 genes. Genes with more than seven transcription factor binding sites were screened. Then, these genes and upregulated genes in colorectal cancer were intersected. Finally, survival analysis and regulatory network analysis were conducted on the screened genes. This allowed clarification of the relationship between these genes and the survival of patients with colorectal cancer and the signaling network involving these genes in the cell signal transduction network of colorectal cancer. Through the above analysis, six upregulated genes in colorectal cancer related to the survival of colorectal cancer patients and highly regulated by transcription factors were selected, namely *STX2*, *PODXL*, *KLK6*, *GRB10*, *EHBP1* and *CREB5*. These genes are involved in signal regulatory networks related to colorectal cancer metastasis-related signaling pathways. Therefore, the survival of patients with colorectal cancer is closely correlated with colorectal cancer metastasis. The six survival-related genes affect the survival of patients by regulating colorectal cancer metastasis-associated signaling pathways.

survival time, colorectal cancer, transcription factor, bioinformatics

Citation: Qi L, Ding YQ. Screening and regulatory network analysis of survival-related genes of patients with colorectal cancer. Sci China Life Sci, 2014, 57: 526−531, doi: 10.1007/s11427-014-4650-1

The survival time of patients with colorectal cancer is influenced by many factors. During the actual study, changes in the expression of some genes in colorectal cancer cells affected the survival time and prognosis of patients with colorectal cancer. Therefore, clarifying the molecular mechanisms of changes in the survival time of colorectal cancer patients has important significance for guiding the treatment of colorectal cancer and evaluating prognosis. In a previous study, survival time-related genes of colorectal cancer patients were screened from the expression profiling data GSE17538 in the GEO database of NCBI using significance analysis software SAM3.01 [1]. A total of 235 survival time-related genes of colorectal cancer patients were obtained, and KEGG pathway enrichment analysis showed that these genes were closely related to colorectal cancer metastasis. In this study, further analysis was conducted primarily on the 235 survival time-related genes, and relatively important genes were screened by regulatory network analysis on interactions of the proteins encoded by these genes. The effects of these genes on the survival time of patients with colorectal cancer were illustrated.

*Corresponding author (email: dyq@fimmu.com)

1 Materials and methods

First, enrichment analysis was conducted on transcription factor binding sites in the transcription regulatory regions of the 235 genes described above using the Transcription Factor Matrix Explorer (TFM-Explorer) tool [2]. Some transcription factor binding sites (TFBS) existed in multiple genes simultaneously, indicating there were more survival-related genes regulated by these transcription factors. Sequencing was conducted according to scores to screen and obtain the set of this type of transcription factor binding sites. The transcription regulatory regions of many genes contain several of the transcription factor binding sites described above, and genes with more than seven transcription factor binding sites were selected.

Because the expression of these genes in colorectal cancer is unknown, genes that are upregulated in colorectal cancer will be of more significance to the survival time of patients. Therefore, upregulated gene sets in colorectal cancer needed to be used to filter screened genes above. GSE21815 and GSE21510 in the GEO database were used to screen upregulated gene sets in colorectal cancer. The previous analysis showed that the enrichment outcome of 235 survival-related genes was correlated with colorectal cancer metastasis, therefore data on nine cases of normal colorectum and data on 49 cases of colorectal cancer in lymph node metastasis were selected from GSE21815 for comparison and screening. Additionally, data on 25 cases of normal colorectum and data on 39 cases of colorectal cancer in lymph node metastasis were selected from GSE21510 for comparison and screening. Differentially expressed genes were screened using Genespring 11.5. The false discovery rate was assessed applying the Benjamini-Hochberg method. The P value of screened genes was adjusted to less than 0.01, and genes in which changes in expression in colorectal cancer were more than two times greater than those in normal colons were selected.

Screened genes with more than seven transcription factor binding sites and upregulated differentially expressed genes screened by expression profiles of the above two groups were intersected, and then the two groups of filtered genes were intersected (Figure 1). The regulatory network analysis was conducted on this set of screened genes using STRING [3]. Protein nodes in the network were extracted and pathway enrichment analysis was performed through the KEGG database, and proteins involved in the network were clarified.

2 Results

2.1 Enrichment results of transcription factor binding sites of survival-related genes

To determine the transcription factor regulation of the 235 survival-related genes, enrichment analysis was conducted on transcription factor binding sites in the transcription regulatory regions using TFM-Explorer. Twenty-five transcription factor binding sites with higher frequency of occurrences were screened from the 235 genes. Most transcription factor binding sites were located between nucleotides 20 and 240 upstream of the transcription initiation site. Among the 25 enriched transcription factor binding sites, SOX2, NFIC and TBP presented a repetition based on different locations, so there were a total of 22 types of transcription factor binding sites: SP1, EGR1, ELK4, EN1, ESRRB, ETS1, FOXC1, FOXD1, FOXQ1, GATA2, KLF4, NFIC, NFYA, PDX1, RELA, RREB1, SOX2, SOX5, SPIB, SRF, T and TBP. This indicates that these transcription factors might regulate the part of 235 survival-related genes, so these transcription factors have an important significance in the regulation of the survival time of patients with colorectal cancer.

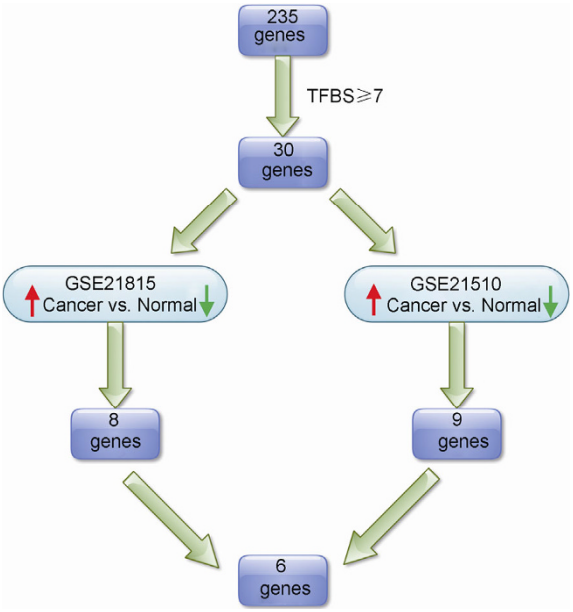

Figure 1 Screening process of key genes related to survival time.

2.2 Screening results of key genes related to survival

Because the transcription regulation of different genes varies, the number of upstream transcription factor binding sites of genes also varies. The greater the number of transcription factor binding sites, the more transcription factors that regulate the gene, so the more complex the regulatory mechanisms that govern the gene expression, and therefore, the greater impact of the gene on life activities. Multiple transcription factor binding sites usually constitute *cis*-regulatory modules (CRM). Studies have shown that CRM-enriched regions are associated with growth-related transcription factors, and growth-related transcription factors also play an important role in the occurrence and de-

velopment of tumors [4]. Therefore, genes with more binding sites for the above 25 transcription factors may be more important in the development of colorectal cancer, as well as the survival time of patients with colorectal cancer. The survival-related genes containing the most transcription factor binding sites were *CREB5* and *LOC401317*. These two genes contain binding sites for 13 of the 25 transcription factors. Genes *ETV1*, *MEIS2* and *PALLD* contain binding sites for 12 transcription factors, genes *EHBP1*, *MITF* and *TNIK* contain binding sites for 11 transcription factors, and gene *TGFB1I1* contains binding sites for 10 transcription factors. Considering the number of screened genes and the number of transcription factors, 30 genes containing more than seven transcription factor binding sites were selected. Upregulated genes in tumors generally have a positive impact on the occurrence and development of tumors. If the upregulated genes are negatively correlated with the survival time of patients, the survival rate of patients will be significantly decreased with the increase of gene expression. Therefore, the survival time of patients could be assessed according to expression of these genes. Second, these genes showed positive expressions; their expression was increased in colorectal cancer and decreased in normal colorectal tissues, therefore the genes were more conducive in clinical diagnosis. During cell signal transduction, the upregulated genes could be involved in a cascade of upregulated signals, thereby affecting the biological behavior of cells. For example, upregulated genes were adopted by GSEA [5] and other enrichment tools for the enrichment analysis of different biological phenotypes. Therefore, in this study, further screening needed to be conducted on genes with more than seven transcription factor binding sites. Upregulated gene sets in colorectal cancer were screened using data GSE21815 and GSE21510 in the GEO database, and then the gene set was used to intersect with the above 30 genes. Eight genes were obtained after intersection with upregulated gene sets of data GSE21815, including *TGFB1I1*, *STX2*, *PODXL*, *KLK6*, *GRB10*, *EHBP1*, *CREB5* and *ARHGAP5*. Nine genes were obtained after the intersection with upregulated gene sets of data GSE21510, including *ZNF83*, *STX2*, *PODXL*, *KLK6*, *GRB10*, *ETV1*, *CREB5*, *CALD1* and *EHBP1*. Two groups of intersected genes were intersected again, and six survival time-related genes with more transcription factor binding sites upregulated in colorectal cancer were screened, including *STX2*, *PODXL*, *KLK6*, *GRB10*, *EHBP1* and *CREB5*.

2.3 Results of survival analysis on survival-related key genes

To further clarify the survival curves of the above six genes, a survival analysis was conducted on these genes using expression profiling data GSE17536 in the GEO database [6]. Data GSE17536 and data GSE17538 used in the previous study showed a containment relationship. Data GSE17536 contain expression profiling data of 177 colorectal cancer tissues with information on patient survival time. The chip platform was Affymetrix Human Genome U133 Plus 2.0 Array. Expression values of the above six genes of 177 cases were extracted from the expression profiling data. Expression values of the six genes were sorted from small to large, and genes were divided into low, medium and high expression groups. Thus, each group had expression values of 59 cases, and expression values of the three groups showed an increasing relationship. Then, the survival rate of the three groups was analyzed using the Kaplan-Meier survival analysis method to obtain the survival curve (Figure 2). Based on changes in gene expression values, *CREB5* (P=0.0036), *EHBP1* (P=0.0029), *GRB10* (P=0.0396), *PODXL* (P=0.0007), *STX2* (P=0.0018) showed a relatively apparent difference in the survival rate, all with P values less than 0.05. Though the P value of *KLK6* (P=0.1069) was greater than 0.05, the median survival time of patients with low gene expression was significantly higher than those with high gene expression. Reports in the literature support this result, indicating that *KLK6* also correlated with patient survival. Almost all results of the survival curve of these six genes showed the survival rate was high in patients with low expressions of these genes, and low in patients with high expressions of these genes, indicating these genes had a positive impact on the occurrence and development of colorectal cancer. It could be seen from the survival curve that the six genes had inconsistent impacts on the tumor; for example, the survival rate was relatively high when *STX2* and *CREB5* were at a low level of expression. However, the survival curve of patients with a medium level of gene expression was close to that of patients with a high level of gene expression, and both survival rates were relatively low, indicating that these genes had greater impacts on the development of colorectal cancer and thus the survival time of patients.

2.4 Results of regulatory network analysis on key survival-related genes

The six screened survival-related genes all exhibit specific and independent functions, but their functions are mutually coordinated and intermeshed. To further clarify the interactions among the six survival-related genes in the cell signaling network of colorectal cancer, the intermeshing of the interacting proteins of the six genes was analyzed using the STRING tool. To improve the reliability of interacting proteins, experiment-based protein connection was selected from parameter. Regulatory networks among proteins were complex, and an excessive number of interacting proteins would hinder our ability to obtain conclusive observations, so the least number of proteins possible were connected to the six proteins to construct a protein interaction network. Therefore, the number of proteins was controlled to 30 in the network (Figure 3). Then, KEGG pathway enrichment

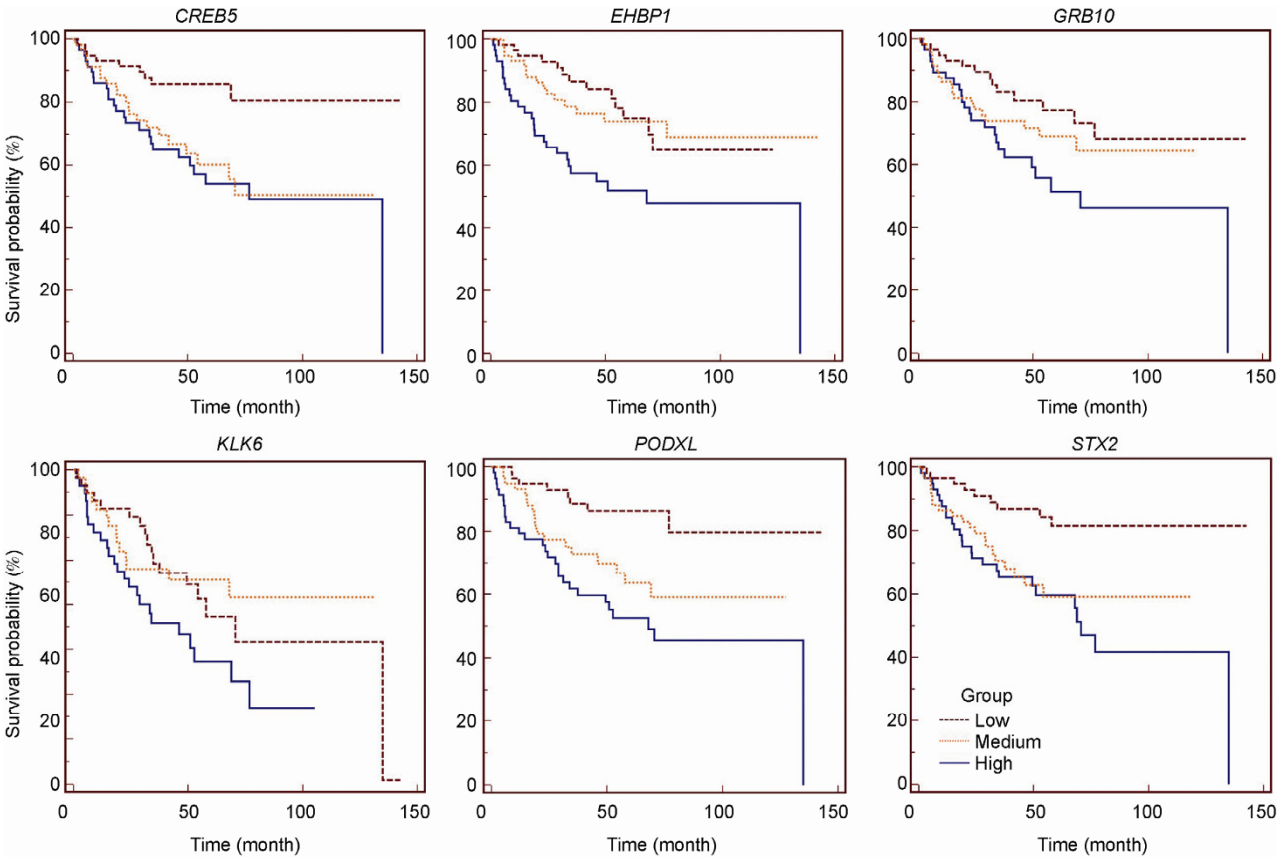

Figure 2 Survival analysis diagram of six survival-related genes.

was conducted on the above 30 protein nodes in the network using the DAVID tool [7] (Table 1). Results showed that proteins in the network were mainly involved in the related signal regulation of colorectal cancer, focal adhesion, ERbB signaling pathways and tumors. GO analysis showed that most proteins in the network were localized on the cell membrane and were also associated with cell secretion, indicating that proteins in the network were closely related to cell adhesion. Therefore, the protein interaction network composed of six screened key genes related to the survival of colorectal cancer patients was not only associated with colorectal cancer, but also more related to colorectal cancer metastasis, indicating the six survival-related genes affected the survival time of colorectal cancer patients by regulating colorectal cancer metastasis.

3 Discussion

Colorectal cancer metastasis determines patient survival time. The occurrence of metastasis is related to the treatment and prognosis of patients. In this study, the relationship between differences in the expression of these genes and patient survival time was clarified by screening key genes related to the survival of colorectal cancer patients. Through the regulatory network analysis of protein interactions, these genes were found to play a role by regulating metastasis-related signaling pathways, further confirming the relationship between colorectal cancer metastasis and patient survival time. As seen from the survival curve of the six screened genes, the survival rate of patients is significantly decreased with the increase of gene expression, indicating that high expression of these genes promotes the development of colorectal cancer. Furthermore, survival curves overlap when *STX2* and *CREB5* genes are at medium and high expression levels, and they show a clear difference when genes showed low expression, indicating medium expression of these two genes can achieve the same degree of impact as high expression on patient survival rate, thus indicating that the tumor promoting effect of the two genes is more intense. It has been reported that *KLK6* [8–11] and *PODXL* [12–14] genes are highly expressed in colorectal cancer and correlate with patient prognosis. The survival rate will decrease if their expressions are elevated, which is consistent with the results of this study and further illustrates the reliability of the present findings.

The survival time is closely related to the treatment and prognosis of colorectal cancer patients. Though many genes related to the survival time of colorectal cancer patients were found in the previous study, these genes affect the

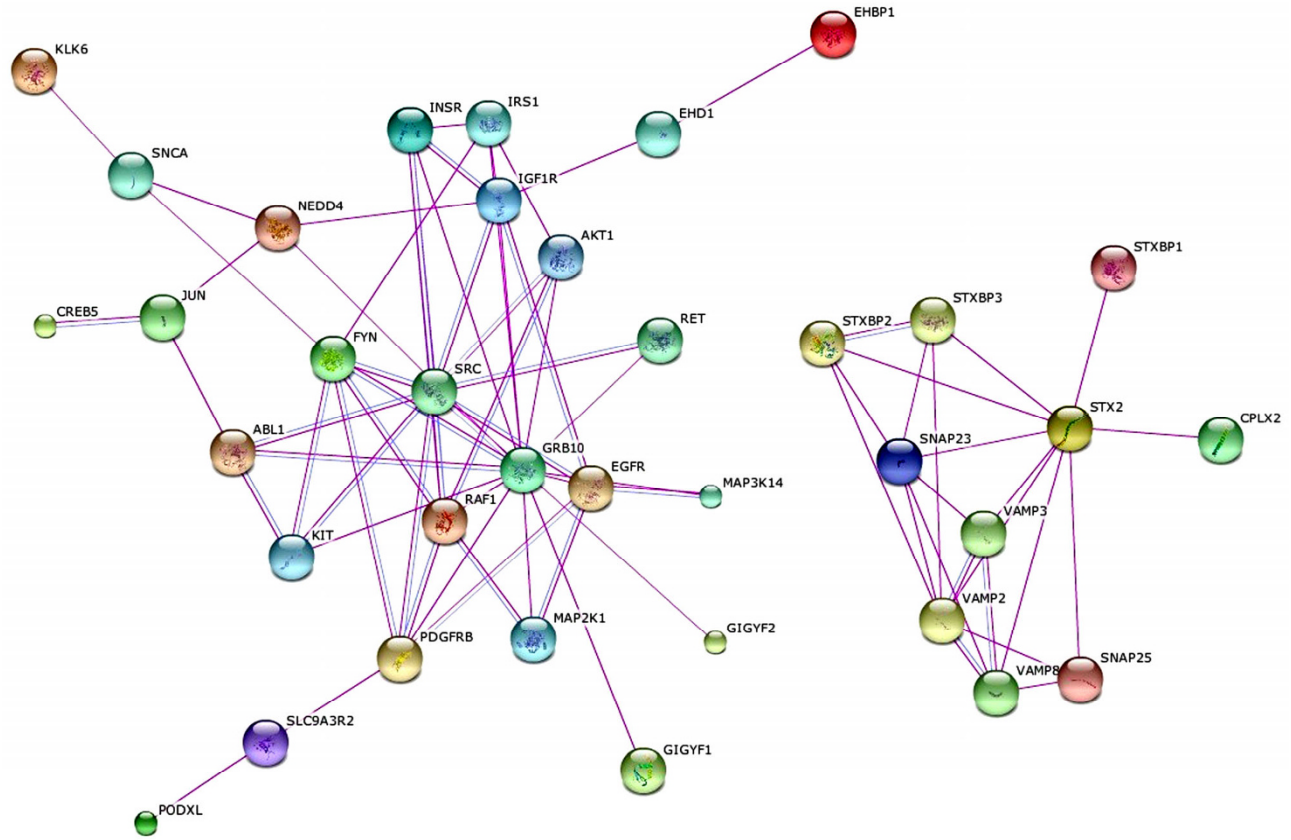

Figure 3 Interaction network graphs of six survival-related genes.

Table 1 Results of KEGG pathway enrichment of 30 interacting protein nodes in the network

Category	Term	Count	*P*-value
KEGG_PATHWAY	hsa05210:Colorectal cancer	7	1.37×10^{-6}
KEGG_PATHWAY	hsa04510:Focal adhesion	9	1.53×10^{-6}
KEGG_PATHWAY	hsa04012:ErbB signaling pathway	7	1.69×10^{-6}
KEGG_PATHWAY	hsa05200:Pathways in cancer	10	6.32×10^{-6}
KEGG_PATHWAY	hsa05214:Glioma	6	7.04×10^{-6}
KEGG_PATHWAY	hsa05218:Melanoma	6	1.27×10^{-5}
KEGG_PATHWAY	hsa04130:SNARE interactions in vesicular transport	5	2.12×10^{-5}
KEGG_PATHWAY	hsa05215:Prostate cancer	6	3.85×10^{-5}
KEGG_PATHWAY	hsa04660:T cell receptor signaling pathway	6	9.77×10^{-5}
KEGG_PATHWAY	hsa04144:Endocytosis	7	1.25×10^{-4}
KEGG_PATHWAY	hsa04722:Neurotrophin signaling pathway	6	1.88×10^{-4}
KEGG_PATHWAY	hsa04520:Adherens junction	5	3.46×10^{-4}
KEGG_PATHWAY	hsa04540:Gap junction	5	6.02×10^{-4}
KEGG_PATHWAY	hsa04912:GnRH signaling pathway	5	8.67×10^{-4}
KEGG_PATHWAY	hsa04010:MAPK signaling pathway	7	9.37×10^{-4}

growth and migration of tumor cells by regulating other related proteins rather than individually affecting patient survival, and many of the regulated genes are associated with tumor metastasis. The more regulatory mechanisms the genes are involved in, the more critical their functions. Thus, by screening genes containing more transcription factor binding sites, genes playing a key role in patient survival time can be better identified, and changes in the expression of these genes will greatly affect the development of tumor cells and have an important impact on the survival time and prognosis of patients. In this study, six genes closely related to the survival time of patients with colorectal cancer were screened. By detecting the expression of these genes, the prognosis and survival time of patients with colorectal cancer may be able to be assessed to determine the best time for treatment.

This work was supported by the Joint Funds of the National Natural Science Foundation of China (U1201226) and the National Natural Science Foundation of China (30670967).

1 Qi L, Yuan L, Wu P, Ye YP, Ding YQ. Screen molecular signatures related to colorectal cancer metastasis (In Chinese). Prog Mod Biomed, 2012, 12: 5225–5229
2 Tonon L, Touzet H, Varre JS. TFM-Explorer: mining *cis*-regulatory regions in genomes. Nucleic Acids Res, 2010, 38: W286–292
3 Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res, 2013, 41: D808–815
4 Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res, 2006, 16: 656–668
5 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, 102: 15545–15550
6 Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res, 2005, 33: D562–566
7 Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res, 2007, 35: W169–175
8 Ogawa K, Utsunomiya T, Mimori K, Tanaka F, Inoue H, Nagahara H, Murayama S, Mori M. Clinical significance of human kallikrein gene 6 messenger RNA expression in colorectal cancer. Clin Cancer Res, 2005, 11: 2889–2893
9 Kim JT, Song EY, Chung KS, Kang MA, Kim JW, Kim SJ, Yeom YI, Kim JH, Kim KH, Lee HG. Up-regulation and clinical significance of serine protease kallikrein 6 in colon cancer. Cancer, 2011, 117: 2608–2619
10 Petraki C, Dubinski W, Scorilas A, Saleh C, Pasic MD, Komborozos V, Khalil B, Gabril MY, Streutker C, Diamandis EP, Yousef GM. Evaluation and prognostic significance of human tissue kallikrein-related peptidase 6 (KLK6) in colorectal cancer. Pathol Res Pract, 2012, 208: 104–108
11 Ohlsson L, Lindmark G, Israelsson A, Palmqvist R, Oberg A, Hammarstrom ML, Hammarstrom S. Lymph node tissue kallikrein-related peptidase 6 mRNA: a progression marker for colorectal cancer. Br J Cancer, 2012, 107: 150–157
12 Barderas R, Mendes M, Torres S, Bartolome RA, Lopez-Lucendo M, Villar-Vazquez R, Pelaez-Garcia A, Fuente E, Bonilla F, Casal JI. In-depth characterization of the secretome of colorectal cancer metastatic cells identifies key proteins in cell adhesion, migration, and invasion. Mol Cell Proteomics, 2013, 12: 1602–1620
13 Larsson A, Johansson ME, Wangefjord S, Gaber A, Nodin B, Kucharzewska P, Welinder C, Belting M, Eberhard J, Johnsson A, Uhlen M, Jirstrom K. Overexpression of podocalyxin-like protein is an independent factor of poor prognosis in colorectal cancer. Br J Cancer, 2011, 105: 666–672
14 Larsson A, Fridberg M, Gaber A, Nodin B, Leveen P, Jonsson G, Uhlen M, Birgisson H, Jirstrom K. Validation of podocalyxin-like protein as a biomarker of poor prognosis in colorectal cancer. BMC Cancer, 2012, 12: 282