doi:10.4149/neo_2015_090

Identification of genes in ulcerative colitis associated colorectal cancer based on centrality analysis of co-expression network

J. ZHU[1,‡,*], C. LI[2,‡], W. JI[1]

[1]Department of General Surgery, Jinan Military General Hospital, No.25 Shifan Road, Jinan 250031, Shandong Province, People´s Republic of China; [2]The Department of Geriatrics Cardiology, Jinan Military General Hospital, No.25 Shifan Road, Shandong Province, Jinan 250031, People´s Republic of China

**Correspondence: jinming_zhu@yeah.net*
‡Contributed equally to this work.

Received January 8, 2015 / Accepted May 25, 2015

PreviousColorectal cancer (CRC) is a well-recognized complication of Ulcerative colitis (UC) and patients with UC have a higher incidence of CRC than the general population. Early detection and mechanism of colitis-associated colorectal cancer (CAC) is still challenging. The aim of present study is to identify genes associated with CAC by centrality analysis of co-expression networks. Co-expression networks of CRC and UC were constructed by empirical Bayes approach based on top 200 gene signatures which identified by the model of genome-wide relative significance and genome-wide global significance across multiple datasets. Centrality of degree, stress centrality, betweenness centrality and closeness centrality of co-expression networks were selected to explore hub genes presented in CRC and UC. Validation of mRNA expression in CRC patients was conducted by real-time quantitative Polymerase Chain Reaction (qPCR). Pathway analysis was conducted based on Kyoto Encyclopedia of Genes and Genomes database. We found 21 common genes, such as *SLC4A4* and *AQP8*, both existed in CRC and UC top 200 genes. By accessing centralities analyses of co-expression networks, *HPGD* and *AQP8* were common hub genes in CRC and UC, and various centralities analyses of the same gene were not consistent. Patients with alteration of *AQP8* have significantly reduced the survival rate according to real-time qPCR results. Our study displayed genes associated with CAC (*AQP8* and *HPGD*), and they might be reliable biomarkers for early detection and therapies of CAC.

Key words: ulcerative colitis, colorectal cancer, centrality, gene

Colorectal cancer (CRC) usually developed from ulcerative colitis (UC), and is one of the commonest malignant tumors with relatively poor prognosis [1, 2]. An increased risk of colitis-associated CRC (CAC) compared to individuals without UC has been presented [3]. The increased incidence occurs predominantly in patients with longstanding extensive colitis [4]. Although CAC accounts only for 1% of all cases of CRC seen in the general population, it is a serious sequel of the disease and accounts for one sixth of all deaths in UC patients [5].

Recently, identifying independent effects of individual gene in multiple existing genome, association has been utilized to account for mechanism of CRC, especially CAC [6, 7]. Hiromu Suzuki et al evaluated a group of genes that were preferentially hypermethylated in CRC, such as *SFRP1* [8]. In addition, *p14* and *COX-2* were identified as potential biomarkers for early detection of CAC [9, 10]. However, traditional gene research ignores that genes are not only encoded as individual genes or proteins, but also as sub-networks of interacting proteins within a larger interaction network in the human genome [11]. As a result, much of the mechanism of human diseases such as CAC remains unexplained.

Unveiling CAC mechanism still has remained a major challenge despite numbers of researches have been conducted. Inconsistent results have been presented due to multiple sources of problems, including small sample size, measurement error, and different statistical methods. The overlap is very low for the most significantly dys-regulated genes across multiple studies [12]. Network-based approaches especially co-expression network offer effective means to at least partially solve this challenge with providing potential malignancy diagnostic molecular signatures and connecting them together.

The aim of present study is to identify genes associated with CAC by centrality analysis of co-expression networks. We constructed co-expression networks utilizing empirical Bayes (EB) approach via linking gene signatures which is evaluated by genome-wide global significance (GWGS) method. Besides, centrality of degree and three kinds of centralities (stress, betweenness and closeness centrality) on the basis of co-expression networks were analyzed to explore hub genes existed in UC and CRC. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed to find functional relevance of selected gene signatures based on expression analysis systematic explored (EASE) test. Finally, real-time quantitative Polymerase Chain Reaction (qPCR) was applied to validate genes mRNA expressions of CRC and patients survival status. As the result, this research might provide the promising gene signatures for therapy of CAC.

Materials and methods

Data selection. We firstly explored UC and CRC related data in Gene Expression Omnibus (GEO) and Array Express (AE) database, then screened these data with similar conditions (such as possessing normal controls, total RNA obtained from intestinal biopsies and clear sample descriptions), and finally six microarray expression profiles (GSE36807 [13], GSE38713 [14], GSE6731 [15], GSE4183 [16], GSE41258 [17] and E-MTAB-57 [18]) were selected. There were total 90 UC patients and 24 normal controls for UC analysis, while a total of 350 CRC patients and 140 normal controls were used. The characteristics of data were shown in S1.

Data preprocess. For each dataset, we applied standard methods to control quality of gene microarray probe-level data [19]. Briefly, in order to eliminate the influence of nonspecific hybridization, background correction was applied by robust multi-array average (RMA) method [20]. The observed Perfect match (PM) probes were modeled as the sum of a normal noise component *N* (Normal with mean μ and variance σ^2) and an exponential signal component *S* (exponential with mean α). To avoid any possibility of negatives, the normal was truncated at zero. Given we had *O* the observed intensity, this then leaded to an adjustment.

$$E(s \mid O=o) = a + b\frac{\phi(\frac{a}{b}) - \phi(\frac{o-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{o-a}{b}) - 1}$$

Where a=s-μ-$\sigma^2\alpha$ and b=σ. Note that ∅ and Φ were the standard normal distribution density and distribution functions respectively. Mismatch (MM) probe intensities were not corrected by the routine.

Normalization was performed through quantiles based algorithm [21]. It was a specific case of the transformationx_i' = $F^{-1}(G(x_i))$, where we estimated *G* by the empirical distribution of each array and *F* using the empirical distribution of the averaged sample quantiles. Using "mas" method to carry out PM/MM correction [19]. An ideal mismatch was subtracted from PM. The Ideal MM would always be less than the corresponding PM and thus we could safely subtract it without risk of negative values.

The summarization method was "medianpolish" [20]. A multichip linear model was fit to data from each probe set. In particular for a probe set *k* with i=1, …, I_k probes and data from j=1,…, *J* arrays we fitted the following model

$$\log_2(\mathrm{PM}_{ij}^k) = \alpha_i^k + \beta_j^k + \varepsilon_{ij}^k$$

Where α_i was a probe effect and β_j was the $\log_2$ expression value.

Detecting of gene signatures. The gene signatures were screened by a model: GWRS and GWGS [22]. The value of GWGS was utilized to integrate independent microarrays, a gene with large value was considered to be globally significant across multiple studies. In current research, gene signatures were identified by two steps. First, the GWRS of *i*-th gene in the *j*-th dataset was measured by $s_{ij} = -2\log(\frac{r_{ij}}{m})$. The number of datasets was denoted by *n*, the number of unique genes across *n* datasets was denoted by *m*; r_{ij} (i= 1–m, j = 1–n) was the rank number of *i*-th gene in the *j*-th study. When a gene was mapped to multiple probe-sets, the maximum value was given to indicate the expression of the probe-set. The gene would be removed if it was absent for one dataset. The degree of differential expression of genes was measured by fold-change. We assigned a rank number for each gene according to their differential expression.

Second, GWGS of the genes were measured by $s_j^r = \sum_{j=1}^{n} \omega_j s_{ij}$. The ω_j represented the relative weight of the *j*-th dataset. The value of weight could be assigned based on the data quality of the *j*-th datasets, and the value of ω_j could also be used to reflect the differential importance of biopsy versus cell line samples that biological scientists may wish to take into account. We treated all the dataset equally, thus the weight of each datasets was set equally to be 1/*n* for *j* = 1–*n*. We also selected only the top 200 genes from the full gene list for further analysis (i.e. selected genes with the greatest s^r value) by empirical evaluation of the classification performance.

Co-expression network construction. A multitude of methods have been developed for co-expression analysis to identify differentially co-expressed (DC) gene, but they are often prone to false discoveries under the conditions of large cardinality of the space to be interrogated [23]. Here, an effective approach of EB framework was conducted which provided an false discovery rate (FDR) controlled list of interesting pairs along with pair-specific posterior probabilities [24]. The identification of DC gene pairs was processed at the following steps: three inputs of matrix X, the conditions array and the pattern object were required. The expression values in an m-by-n matrix of X (where m indicated the number of genes/probes under consideration, n indicated the total number of microarrays over all conditions) were normalized with background normalization and median correction and were generally represented on the $\log_2$ scale. The members of the conditions array with length n took values in $1,\ldots\ldots, K$ (K indicated the total number of conditions).

It was used to define the EC/DC classes with an ebarraysPatterns object based on the unique values in the conditions array. Intra-group correlations for all $p=m^*(m-1)/2$ gene pairs from X and the conditions array were calculated using bi-weight mid-correlation through the function makeMyD. The p-by-K of D matrix with correlations was obtained. Mclust algorithm [25] was used to initialize the hyper parameters through the initializeHP function to find the component Normal mixture model which could best fit the empirical distribution of correlations. The values of the component in Normal mixture model with component means, standard deviations and weights would be used to initialize the expectation maximization (EM) algorithm [26]. The three functions of the 'full' version, the 'one-step' version and the 'zero-step' version represented different flavors of the modified EM approach. In this step, the initial estimates of the hyper parameters rather than the 'zero-step' version were used to generate posterior probabilities of DC. After the EM computations were finished with the selected function, the prior diagnostic function for the prior predictive distribution was used to check how well the model chosen by the EM fitted the data. Finally, the crit.fun function was used to provide a soft threshold with controlling the posterior probabilities of DC in order to identify particular types of DC gene pairs. Here, DC genes were distinguished from gene pairs having invariant expression with controlling the posterior expected FDR at 0.05 and the co-expression network was constructed to represent the correlation between each pair of genes.

Centralities analysis of the co-expression network. Many studies demonstrate the presence of strong correlations between the co-expression network structure and the functional role of its protein/gene constituents [22-23]. In order to understand the functionality of complex systems of gene signatures, we characterized the biological importance of genes based on the co-expression network using indices of topological centrality. Centralities related to local (degree) scale, and global (stress centrality, betweenness centrality and closeness centrality) scale which were used to describe the importance of nodes were analyzed.

Degree centrality. Degree quantifies the local topology of each gene, by summing up the number of its adjacent genes [24]. It gives a simple count of the number of interactions of a given node. The genes at the top of degree distribution (>=95% quantile) in the significantly perturbed networks were defined as hub genes. The degree $C_D(v)$ of a node v is defined as

$$C_D(v) = \sum_j a_{vj}$$

Stress centrality. Stress centrality, a node centrality index, is considered by the number of nodes in the shortest path between two nodes. To calculate the stress ($Cstr\ (v)$) of a node v, all shortest paths in a graph G are calculated and then the number of shortest paths passing through v is counted. A "stressed" node is a node traversed by a high number of shortest paths. σst is the total number of shortest paths from node s to node t and σst is the number of those paths that pass through v. $C_{str}(v)$ is calculated as following:

$$C_{str}(v) = \sum_{s\neq v\in N}\sum_{t\neq v\in N}\sigma_{st}(v)$$

Betweenness centrality. Betweenness centrality [25] is another topological metric in graphs for determining how the neighbors of a node are interconnected. It is considered the ratio of the node in the shortest path between two other nodes. The betweenness centrality of a node v is given by the expression:

$$C_B(v) = \sum_{s\neq v\neq t\in N}\frac{\sigma_{st}(v)}{\sigma_{st}}$$

Betweenness centrality of a node scales with the number of pairs of nodes as implied by the summation indices. Therefore the calculation may be rescaled by dividing through by the number of pairs of nodes not including v, so that $C_B(v) \in [0,1]$. σst is the total number of shortest paths from node s to node t and $\sigma st\ (v)$ is the number of those paths that pass through v.

Closeness centrality. Closeness centrality is a measure of the average length of the shortest paths to access all other proteins in the network[27]. The larger the value, the more central is the protein. The closeness centrality, $Cc(v)$ was calculated for every functional category taking into consideration, all of the shortest path for each node. $Cc(v)$ of node n is defined as the reciprocal of the average shortest path length and is computed as follows:

$$C_C(v) = \frac{1}{\sum_{t\in N} d_G(v,t)}$$

Where $dG\ (s,\ t)$ represents the length of the shortest path between two nodes s and t in graph G, which is the sum of the weights of all edges on this shortest path. Meanwhile, $dG\ (s,$ s) = 0, $dG\ (s,\ t) = dG\ (t,\ s)$ in undirected graph.

Pathway enrichment analyses. The Database for Annotation, Visualization, and Integrated Discovery (DAVID) for KEGG pathway enrichment analysis were carried out to further investigate the biological functions of Top 200 genes [28]. KEGG pathways with P value < 0.05 were chosen based on EASE test applied in DAVID. EASE analysis of the regulated genes indicated molecular functions and biological processes unique to each category [29]. The EASE score was used to detected the significant categories. In both of the functional and pathway enrichment analysis, the threshold of minimum number of genes for the corresponding term >2 were considered significant for a category.

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Where *n* was the number of background genes; *a*' was the gene number of one gene set in the gene lists; *a*' + *b* was the number of genes in the gene list including at least one gene set; *a*' + *c* was the gene number of one gene list in the background genes; *a*' was replaced with *a*=*a*'-1.

Real-time qPCR. Samples of 32 colitis-associated colorectal cancer (CAC) patients were obtained from colon surgery. Tissue samples were originated from open tumor resection, whose molecular genetics evaluation was exclusively done in tissue samples in the direct vicinity of samples showing solid tumor tissue [30], and control sample was normal tissue nearby tumor tissue. The mRNA expressions throughout the solid tumor space, biopsy specimens were harvested from at least two different sites along the chosen trajectory of each tumor in the biopsy group. As the next step, purifications for mRNA were conducted [31]. In order to achieve suitable amounts of mRNA for gene expression analyses, a certain quality of purified RNA of all samples were amplified utilizing the Target Amp-Kit (Epicentre, Madison, Wisconsin, USA) according manufacturer's recommendations.

Real-time qPCR was performed with the Light Cycler 480 instrument (Roche Diagnostics, Mannheim, Germany) using Roche's qPCR Mastermix and highly specific Universal ProbeLibrary assays (Roche Diagnostics). The following primers were used: *AQP8*: forward primer: 5'-TGGCCAAGGCGGTGAGT-3'; reverse primer: 5'-GCTCCTGGACTGTCACAAAGG-3'. *HPGD*: forward primer: 5'-TGGTCAATAA

TGCTGGAGTGA-3'; reverse primer: 5'-GGTTCCACTGATAACAGAAACCA-3'. All assays were designed intron-spanning. The thermal cycler conditions comprised 45 cycles of 95°C for 10 s, 60°C for 30 s, and 72°C for 15 s. Three replicates of the assay within or between runs were performed to assess the reproducibility.

The data were normalized to β-actin reference and relative mRNA expression was calculated with the Relative Quantification Software (Roche Diagnostics). We computed mean (μ) and standard deviation (σ) of individual gene (*AQP8* and *HPGD*) expression values in patients samples. Then, patient samples were divided into two groups: (1) group with differential expression level, samples having expression value larger or smaller than μ+σ or μ-σ, respectively; (2) group with normal expression level, samples with expression between μ-σ and μ+σ. Comparison of survival curves were conducted by log-rank (Mantel-Cox) Test [32].

Results

Identification of gene signatures. After normalizing and preprocessing of microarray expression profiles, for CRC, there were 20109, 12493 and 12493 genes in E-GEOD-4183, E-GEOD-41258 and E-MTAB-57, respectively. For UC, 8631 genes were presented in E-GEOD-6731, E-GEOD-36807 and E-GEOD-38713 both contained 20109 genes. The rank value of GWGS was applied to integrate multiple independent dataset, and a gene with large value was considered to be globally significant studies. We identified top 200 genes between CRC or UC patients and normal controls as gene signatures for further analysis. Moreover, 21 common genes, such as *SLC4A4* and *AQP8* were discovered both presented in top 200 genes of CRC and UC, as shown in Table 1.

Co-expression network analysis. Many genes together play important roles in the accomplishment of a biological function, and highly co-expressed genes participate in similar biological processes and pathways. In fact, functionally related genes are frequently co-expressed across the samples. In this paper, we constructed the co-expression networks for top 200 genes in CRC and UC using EB approach. In CRC co-expression network (Figure 1), there were 1646 edges and 200 nodes, among which *CHGA* with the highest degree (61), next were *CLMN* (59) and *NFE2L3* (49). For co-expression network of UC (Figure 2), 182 genes were mapped and 1355 edges were produced, *NMT2* (70), *PTPN21* (68) and *PPID* (61) possessed much higher degree than other genes.

Centralities analyses of co-expression networks. Centralities could indicate the relevance of a gene as functionally capable to hold communicating nodes together of a node in a biological network. We defined that genes at the top of degree distribution (>=95% quantile) in the significantly perturbed networks were hub genes. In present study, hub genes of co-expression networks in CRC and UC were obtained by analyzing centrality of degree and shown in Figure 3. We could find that *HPGD* and *AQP8* were common hub genes of CRC and UC.

Table 1. Common genes of top 200 genes identified from CRC and UC

Number	Gene	Number	Gene
1	*SLC4A4*	12	*LCN2*
2	*AQP8*	13	*PTN*
3	*CA1*	14	*S100A11*
4	*HPGD*	15	*PYCR1*
5	*CXCL1*	16	*VWF*
6	*NFE2L3*	17	*PLOD3*
7	*TEAD4*	18	*ARHGEF9*
8	*PADI2*	19	*ANK3*
9	*PRKACB*	20	*ABCC1*
10	*ACAT1*	21	*CFI*
11	*NPM1*		

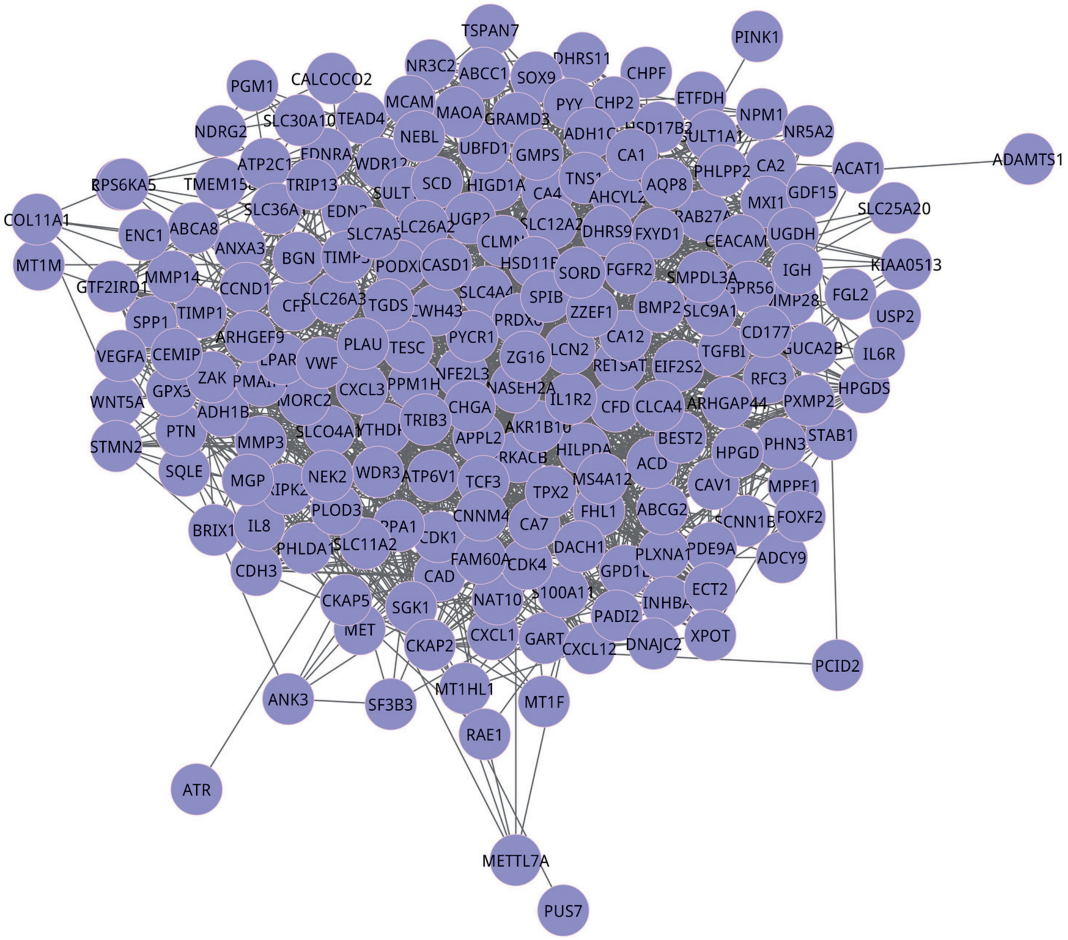


Figure 1. Co-expression network of CRC based on top 200 genes. There were 200 nodes and 1646 edges, where nodes referred to gene signatures and edges between nodes indicated interaction of genes in the network.

Table 2. Top 5% genes of co-expression networks in CRC and UC based on stress centrality and betweenness centrality analysis

Disease	Stress centrality	Betweenness centrality	Closeness centrality
CRC	*PINK1, BMP2, SQLE, MT1F, SLC25A20, TMEM158, FOXF2, XPOT, ATR, LPHN3*	*CLMN, CHGA, NFE2L3, FAM60A, TRIB3,CKAP2, CWH43, ACD, RNASEH2A, IL1R2*	*CHGA, CLMN,NFE2L3, CEMIP, CWH43, FAM60A, TRIB3, APPL2, RETSAT, RNASE-H2A*
UC	*PLEKHO2, GAB1, SPINK2, SLC17A4, HPGD, CFDP1, ZC3H14, PML, P2RY1*	*PTPN21, PPID, NMT2, SMIM8, PRKACB, FMO5, PTGDR, HMGCS2, EAPP*	*NMT2, PTPN21, SMIM8, PPID, FMO5, FTSJ1, ACTA1, YARS, CDC25B*

Table 3. KEGG pathways for CRC and UC

Disease	Terms	P Value	Count
CRC	Nitrogen metabolism	5.28E-04	*CA7, CA12, CA4, CA2, CA1*
	Bladder cancer	5.23E-03	*RPS6KA5, CCND1, IL8, VEGFA, CDK4*
	p53 signaling pathway	2.74E-02	*CDK1, CCND1, ATR, PMAIP1, CDK4*
	Aldosterone-regulated sodium reabsorption	3.14E-02	*SGK1, NR3C2, HSD11B2, SCNN1B*
	Cytokine-cytokine receptor interaction	3.18E-02	*CXCL1, INHBA, IL1R2, BMP2, IL8, CXCL3, MET, VEGFA, IL6R, CXCL12*
UC	Proteasome	1.87E-04	*PSMB5, PSMB10, PSMD14, PSMC5, PSMD12,PSME2, PSMB8*
	Aminoacyl-tRNA biosynthesis	3.83E-02	*WARS, YARS, GARS, MARS*

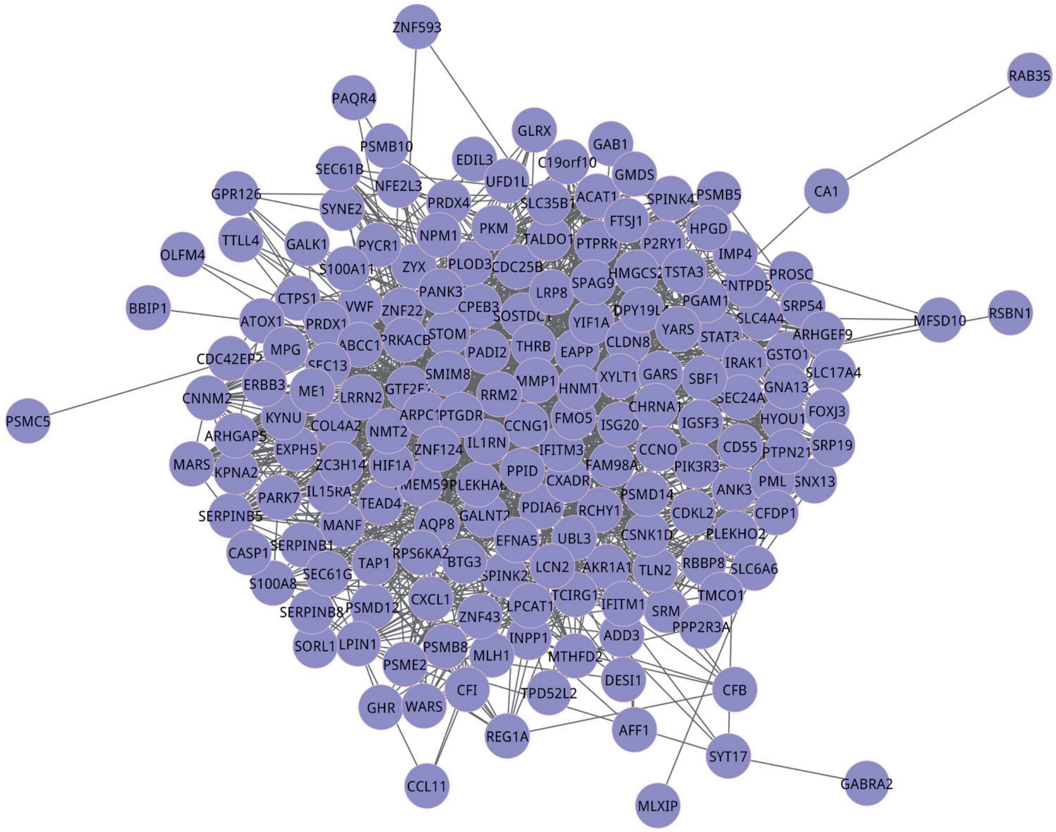


Figure 2. Co-expression network of UC based on top 200 genes. There were 182 nodes and 1355 edges, where nodes referred to gene signatures and edges between nodes indicated interaction of genes in the network.

By assessing stress centrality, betweenness centrality and closeness centrality, centralities of co-expression networks from CRC and UC were obtained, as shown in Table 2. The results revealed that top 5% genes in various centralities analysis of the same gene were not consistent.

Pathway enrichment analysis. We conducted pathway enrichment analysis based on KEGG for CRC and UC, and the results were listed in Table 3. The top 200 genes in CRC was significantly enriched in 5 terms, and the most significant term was nitrogen metabolism (P = 5.28E-04), which contained five genes, such as *CA1*and *CA7*. While for UC, 2 enriched terms were obtained with the threshold of P < 0.05, the most significant one was proteasome (P=1.87E-04).

Clinical outcome. To validate results of network centrality analysis, the expression level of common hub gene (*AQP8* and *HPGD*) was analyzed by real-time qPCR in CAC patients, and we displayed one of the results in supplement material S2. Furthermore, we selected log-rank (Mantel-Cox) test which provided a nonparametric estimate of the survival distribution to compare survival curves of *AQP8* and *HPGD* (Figure 4). The results showed that expressions of *AQP8* were changed in 8 of 32 CAC patients. CAC patients with alteration of *AQP8* (P=0.0387, Chi square=4.273) significantly reduced the survival rate. While for *HPGD,* there was not significantly different in patients with and without alteration (P=0.1814, Chi square=1.786, altered *HPGD* in 5/32 patients).

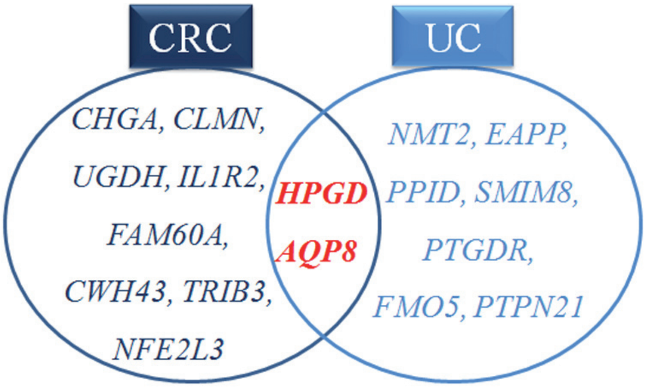


Figure 3. Hub genes of CRC and UC co-expression network based on degree centrality analyses of the network. There were 10 and 9 hub genes of CRC and UC network respectively. *AQP8* and *HPGD* were common hub genes of the networks.

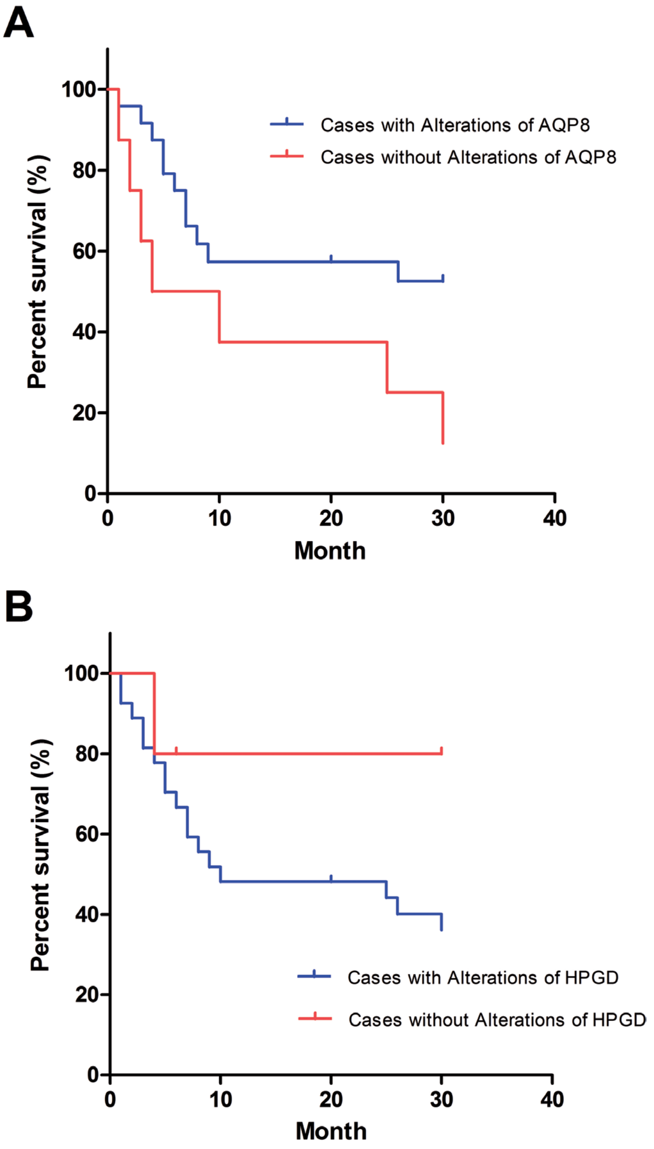


Figure 4. Comparison of survival curves using log-rank (Mantel-Cox) test in 32 patients with CAC. Tumor tissue obtained either by stereotactic biopsy or by open surgery. A: Survival status with altered *AQP8* mRNA expression (P=0.0387); B: Survival status with altered *HPGD* mRNA expression (P=0.1814).

Discussion

In this paper, we identified genes associated with CAC with centralities analysis of co-expression networks in CRC and UC. Co-expression networks for CRC and UC were constructed by EB approach on the basis of top 200 gene evaluated by GWGS method. Degrees and three kinds of centralities (stress, betweenness and closeness centrality) were performed to explore hub genes of CRC and UC. The results showed that 21 common genes, such as *SLC4A4*, *AQP8* and *CA1* presented in top 200 genes of CRC and UC. *HPGD* and *AQP8* were common hub genes of co-expression network in CRC and UC, and various centralities analyses of the same gene were not consistent. Results of real-time qPCR showed that patients with alteration of *AQP8* significantly reduced the survival rate

Patients with UC had an increased risk of developing CAC when compared with the general population [33], and the excess risk was almost entirely confined to patients with long-standing extensive colitis [5]. Important risk factors included primary sclerosing cholangitis [34], family history of CRC [35], whereas the role of other factors, such as age at onset of UC. In present study, 21 common genes were found between CRC and UC. The most significant two genes were *SLC4A4* and *AQP8*, for example, *AQP8* (Aquaporin 8) was a water channel protein and aquaporins were a family of small integral membrane proteins related to major intrinsic protein [36]. The three folds decrease of *AQP8* in UC tissues according to previous research demonstrated that *AQP8* might be involved in the pathogenesis of UC and have a close relationship with miRNA in UC patients [37]. *AQP8* was expressed in all normal colon samples but not, or to a less extent, in the colorectal tumors [38]. Meanwhile Over-expressions of *AQP8* had been implicated in tumorigenesis and proved be a novel prognostic biomarker for CRC patients [39]. Thus we might speculate that some genes contained in UC patients also existed in CRC patients, common genes could declare that if certain genes of UC were inhabited, the risk rate of CRC may be decreased.

Networks as a powerful tool have attracted a great deal of attention to analyze many biological and communication systems. Co-expression network analysis provides an effective way to score and evaluate functionally co-expressed genes across a set of samples from the perspective of systems biology [40]. A key concept of network analysis is node connectivity (centrality), which gives an indication of a gene importance, and a central node (referred to as hub) is one with many connections to other nodes. [41]. In this paper, local (degree) scale, and global (stress centrality, betweenness centrality and closeness centrality) scale were selected to describe the significance of nodes. According to centralities analyses of co-expression network of CRC, *AQP8* and *HPGD* were common hub genes of co-expression network in CRC and UC. In addition, *AQP8* with the highest edge betweenness of 399 and high stress of 3886 was considered the most significant gene signature in CRC regulation. Meanwhile, the mRNA expression of *AQP8* was related to patients' survival status significantly based on the result of overall survival Kaplan-Meier estimation. Therefore *AQP8* might be an important biomarker in the prognosis of CAC.

HPGD, hydroxyprostaglandin dehydrogenase 15-(NAD), is responsible for the metabolism of prostaglandins, which function in a variety of physiologic and cellular processes such as inflammation. *HPGD* had been reported to act as bladder, breast, lung and colorectal tumor suppressor [42, 43]. Previous studies demonstrated that *HPGD* inhibited the development of murine intestinal neoplasias as potent suppressor of the growth of human colon tumor cell lines in immunodeficient mice [43, 44]. These findings deduced that *HPGD* was abolished in various cancers, particularly in human colonic neoplasms, emphasize the oncogenic potential of the prostaglandin synthesis pathway [45]. For instance, Bernd Frank et al revealed that *HPGD* gene variants to be positively associated with CRC risk [46]. Thus we could resume that *HPGD* had a close relationship with inflammation and cancer, and might be potential gene signatures of CAC.

In conclusion, several gene signatures related to CRC and UC were identified, such as *AQP8* and *HPGD*. And they might be potential biomarkers for early detection and therapies of CAC.

Supplementary information is available in the online version of the paper.

Acknowledgements: This research received no specific grants from any funding agency in public, commercial, or not-for-profit sectors.

References

[1] JESS T, RUNGOE C, PEYRIN-BIROULET L, Risk of colorectal cancer in patients with ulcerative colitis: a meta-analysis of population-based cohort studies. Clin Gastroenterol Hepatol 2012; 10: 639–645. http://dx.doi.org/10.1016/j.cgh.2012.01.010

[2] Colitis–Pathophysiology U, Inflammatory bowel disease part I: ulcerative colitis–pathophysiology and conventional and alternative treatment options. Alternative medicine review 2003; 8: 247–283.

[3] ROGLER G, Chronic ulcerative colitis and colorectal cancer. Cancer letters 2013.

[4] KISIEL JB, GARRITY-PARK MM, TAYLOR WR, SMYRK TC, AHLQUIST DA, Methylated Eyes Absent 4 (EYA4) gene promotor in non-neoplastic mucosa of ulcerative colitis patients with colorectal cancer: evidence for a field effect. Inflammatory bowel diseases 2013; 19: 2079–2083. http://dx.doi.org/10.1097/MIB.0b013e31829b3f4d

[5] WATANABE T, KONISHI T, KISHIMOTO J, KOTAKE K, MUTO T, et al., Ulcerative colitis-associated colorectal cancer shows a poorer survival than sporadic colorectal cancer: A nationwide Japanese study. Inflammatory bowel diseases 2011; 17: 802–808. http://dx.doi.org/10.1002/ibd.21365

[6] BARDELLI A, SIENA S, Molecular mechanisms of resistance to cetuximab and panitumumab in colorectal cancer. Journal of Clinical Oncology 2010; 28: 1254–1261. http://dx.doi.org/10.1200/JCO.2009.24.6116

[7] SALEH M, TRINCHIERI G, Innate immune mechanisms of colitis and colitis-associated colorectal cancer. Nature Reviews Immunology 2010; 11: 9–20. http://dx.doi.org/10.1038/nri2891

[8] SUZUKI H, GABRIELSON E, CHEN W, ANBAZHAGAN R, VAN ENGELAND M, et al., A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer. Nature genetics 2002; 31: 141–149. http://dx.doi.org/10.1038/ng892

[9] SATO F, HARPAZ N, SHIBATA D, XU Y, YIN J, et al., Hypermethylation of the p14(ARF) gene in ulcerative colitis-associated colorectal carcinogenesis. Cancer Res 2002; 62: 1148–1151.

[10] WANG D, DUBOIS RN, The role of COX-2 in intestinal inflammation and colorectal cancer. Oncogene 2009; 29: 781–788. http://dx.doi.org/10.1038/onc.2009.421

[11] VINAYAGAM A, ZIRIN J, ROESEL C, HU Y, YILMAZEL B, et al., Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nature methods 2013. http://dx.doi.org/10.1038/nmeth.2733

[12] LIANG D, HAN G, FENG X, SUN J, DUAN Y, et al., Concerted perturbation observed in a hub network in Alzheimer‘s disease. PLoS One 2012; 7: e40498. http://dx.doi.org/10.1371/journal.pone.0040498

[13] MONTERO-MEL NDEZ T, LLOR X, GARC A-PLANELLA E, PERRETTI M, and SU REZ A, Identification of Novel Predictor Classifiers for Inflammatory Bowel Disease by Gene Expression Profiling. PloS one 2013; 8: e76235. http://dx.doi.org/10.1371/journal.pone.0076235

[14] PLANELL N, LOZANO JJ, MORA-BUCH R, MASAMUNT MC, JIMENO M, et al., Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. Gut 2013; 62: 967–976. http://dx.doi.org/10.1136/gutjnl-2012-303333

[15] WU F, DASSOPOULOS T, COPE L, MAITRA A, BRANT SR, et al., Genome-wide gene expression differences in Crohn‘s disease and ulcerative colitis from endoscopic pinch biopsies: insights into distinctive pathogenesis. Inflammatory bowel diseases 2007; 13: 807–821. http://dx.doi.org/10.1002/ibd.20110

[16] GYORFFY B, MOLNAR B, LAGE H, SZALLASI Z, EKLUND AC, Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. PloS one 2009; 4: e5645. http://dx.doi.org/10.1371/journal.pone.0005645

[17] SHEFFER M, BACOLOD MD, ZUK O, GIARDINA SF, PINCAS H, et al., Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proceedings of the National Academy of Sciences 2009; 106: 7131–7136. http://dx.doi.org/10.1073/pnas.0902232106

[18] ANCONA N, MAGLIETTA R, PIEPOLI A, D‘ADDABBO A, COTUGNO R, et al., On the statistical assessment of classifiers using DNA microarray data. BMC bioinformatics 2006; 7: 387. http://dx.doi.org/10.1186/1471-2105-7-387

[19] BOLSTAD B, affy: Built-in Processing Methods. 2013.

[20] IRIZARRY RA, BOLSTAD BM, COLLIN F, COPE LM, HOBBS B, et al., Summaries of Affymetrix GeneChip probe level data. Nucleic acids research 2003; 31: e15-e15. http://dx.doi.org/10.1093/nar/gng015

[21] BOLSTAD BM, IRIZARRY RA, ASTRAND M, SPEED TP, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19: 185–193. http://dx.doi.org/10.1093/bioinformatics/19.2.185

[22] LIU W, PENG Y, TOBIN DJ, A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis. PeerJ 2013; 1: e49. http://dx.doi.org/10.7717/peerj.49

[23] CHOI JK, YU U, YOO OJ, KIM S, Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics 2005; 21: 4348–4355. http://dx.doi.org/10.1093/bioinformatics/bti722

[24] DAWSON JA, YE S, KENDZIORSKI C, R/EBcoexpress: an empirical Bayesian framework for discovering differential co-expression. Bioinformatics 2012; 28: 1939–1940. http://dx.doi.org/10.1093/bioinformatics/bts268

[25] FRALEY C, RAFTERY AE, Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 2002; 97: 611–631. http://dx.doi.org/10.1198/016214502760047131

[26] MOON TK, The expectation-maximization algorithm. Signal processing magazine, IEEE 1996; 13: 47–60. http://dx.doi.org/10.1109/79.543975

[27] WASSERMAN S, Social network analysis: Methods and applications: Cambridge university press, 1994. http://dx.doi.org/10.1017/CBO9780511815478

[28] HUANG DW, SHERMAN BT, LEMPICKI RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2008; 4: 44–57. http://dx.doi.org/10.1038/nprot.2008.211

[29] FORD G, XU Z, GATES A, JIANG J, FORD BD, Expression Analysis Systematic Explorer (EASE) analysis reveals differential gene expression in permanent and transient focal stroke rat models. Brain research 2006; 1071: 226–236. http://dx.doi.org/10.1016/j.brainres.2005.11.090

[30] KRETH S, THON N, EIGENBROD S, LUTZ J, LEDDEROSE C, et al., O6-methylguanine-DNA methyltransferase (MGMT) mRNA expression predicts outcome in malignant glioma independent of MGMT promoter methylation. PLoS One 2011; 6: e17156. http://dx.doi.org/10.1371/journal.pone.0017156

[31] THON N, EIGENBROD S, GRASBON-FRODL EM, RUITER M, MEHRKENS JH, et al., Novel molecular stereotactic biopsy procedures reveal intratumoral homogeneity of loss of heterozygosity of 1p/19q and TP53 mutations in World Health Organization grade II gliomas. Journal of Neuropathology & Experimental Neurology 2009; 68: 1219–1228. http://dx.doi.org/10.1097/NEN.0b013e3181bee1f1

[32] KAPLAN EL, MEIER P, Nonparametric estimation from incomplete observations. Journal of the American statistical association 1958; 53: 457–481. http://dx.doi.org/10.1080/01621459.1958.10501452

[33] GILLEN C, WALMSLEY R, PRIOR P, ANDREWS H, ALLAN R, Ulcerative colitis and Crohn‘s disease: a comparison of the colorectal cancer risk in extensive colitis. Gut 1994; 35: 1590–1592. http://dx.doi.org/10.1136/gut.35.11.1590

[34] LOFTUS E, HAREWOOD G, LOFTUS C, TREMAINE W, HARMSEN W, et al., PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis. Gut 2005; 54: 91–96. http://dx.doi.org/10.1136/gut.2004.046615

[35] ASKLING J, DICKMAN PW, KARL N P, BROSTR M O, LAPIDUS A, et al., Family history as a risk factor for colorectal cancer in inflammatory bowel disease. Gastroenterology 2001; 120: 1356–1362. http://dx.doi.org/10.1053/gast.2001.24052

[36] GARCIA F, KIERBEL A, LAROCCA MC, GRADILONE SA, SPLINTER P, et al., The water channel aquaporin-8 is mainly intracellular in rat hepatocytes, and its plasma membrane insertion is stimulated by cyclic AMP. Journal of Biological Chemistry 2001; 276: 12147–12152. http://dx.doi.org/10.1074/jbc.M009403200

[37] MINMIN P, GANG S, MING-ZHOU G, ZE-WU Q, YUN-SHENG Y, Aquaporin 8 expression is reduced and regulated by microRNAs in patients with ulcerative colitis. Chinese medical journal 2013; 126: 1532–1537.

[38] FISCHER H, STENLING R, RUBIO C, LINDBLOM A, Differential expression of aquaporin 8 in human colonic epithelial cells and colorectal tumors. BMC physiology 2001; 1: 1. http://dx.doi.org/10.1186/1472-6793-1-1

[39] WANG W, LI Q, YANG T, BAI G, LI D, et al., Expression of AQP5 and AQP8 in human colorectal carcinoma and their clinical significance. World J Surg Oncol 2012; 10: 242. http://dx.doi.org/10.1186/1477-7819-10-242

[40] LEE HK, HSU AK, SAJDAK J, QIN J, PAVLIDIS P, Coexpression analysis of human genes across many microarray data sets. Genome research 2004; 14: 1085–1094. http://dx.doi.org/10.1101/gr.1910904

[41] ZHANG B, HORVATH S, A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 2005; 4. http://dx.doi.org/10.2202/1544-6115.1128

[42] MYUNG S-J, RERKO RM, YAN M, PLATZER P, GUDA K, et al., 15-Hydroxyprostaglandin dehydrogenase is an in vivo suppressor of colon tumorigenesis. Proceedings of the National Academy of Sciences 2006; 103: 12098–12102. http://dx.doi.org/10.1073/pnas.0603235103

[43] NA H-K, PARK J-M, LEE HG, LEE H-N, MYUNG S-J, et al., 15-Hydroxyprostaglandin dehydrogenase as a novel molecular target for cancer chemoprevention and therapy. Biochemical pharmacology 2011; 82: 1352–1360. http://dx.doi.org/10.1016/j.bcp.2011.08.005

[44] YAN M, MYUNG S-J, FINK SP, LAWRENCE E, LUTTERBAUGH J, et al., 15-Hydroxyprostaglandin dehydrogenase inactivation as a mechanism of resistance to celecoxib chemoprevention of colon tumors. Proceedings of the National Academy of Sciences 2009; 106: 9409–9413. http://dx.doi.org/10.1073/pnas.0902367106

[45] MANN JR, BACKLUND MG, BUCHANAN FG, DAIKOKU T, HOLLA VR, et al., Repression of prostaglandin dehydrogenase by epidermal growth factor and snail increases prostaglandin E2 and promotes cancer progression. Cancer research 2006; 66: 6649–6656. http://dx.doi.org/10.1158/0008-5472.CAN-06-1787

[46] FRANK B, HOEFT B, HOFFMEISTER M, LINSEISEN J, BREITLING LP, et al., Association of hydroxyprostaglandin dehydrogenase 15-(NAD)(HPGD) variants and colorectal cancer risk. Carcinogenesis 2010bgq231.

doi:10.4149/neo_2015_090

Supplementary Information

Identification of genes in ulcerative colitis associated colorectal cancer based on centrality analysis of co-expression network

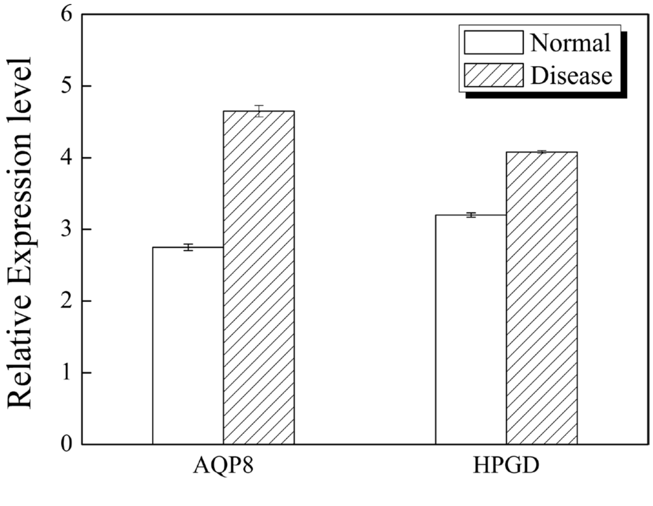
J. ZHU[1,‡,*], C. LI[2,‡], W. JI[1]

[1]Department of General Surgery, Jinan Military General Hospital, No.25 Shifan Road, Jinan 250031, Shandong Province, People´s Republic of China; [2]The Department of Geriatrics Cardiology, Jinan Military General Hospital, No.25 Shifan Road, Shandong Province, Jinan 250031, People´s Republic of China

**Correspondence: jinming_zhu@yeah.net*
‡Contributed equally to this work.

S1 Characteristics of the gene expression profiles

Disease type		UC			CRC		
Accession number		GSE36807	GSE38713	GSE6731	GSE4183	GSE41258	E-MTAB-57
Platform of Affymetrix HG-		U133_Plus_2	U133_Plus_2	U95Av2	U133_Plus_2	U133A	U133A
Total size (Disease/Control)		35 (28/7)	43(30/13)	36(32/4)	53(30/23)	390(290/100)	47(25/22)
Disease	Gender male	16	7	-	13	-	14
	Age, year	-	44.8±10.0	-	68.4±12.9	-	60±14
	Smokers	5	-	-	-	-	-
	Dyslipidemia	1	-	-	9	-	-
Control	Gender male	-	5	-	7	-	12
	Average age, year	-	41.6±12.4	-	40.3±9.9	-	60±28
	Smokers	-	-	-	-	-	-
	Dyslipidemia	-	-	-	9	-	-


Relative Expression level
Normal
Disease
AQP8
HPGD