Published in final edited form as:
Transl Cancer Res. 2013 February 1; 2(1): 6–17. doi:10.3978/j.issn.2218-676X.2012.12.04.

Utilizing *signature*-score to identify oncogenic pathways of cholangiocarcinoma

Tzu-Hung Hsiao[1], **Hung-I Harry Chen**[1], **Jo-Yang Lu**[2], **Pei-Ying Lin**[2], **Charles Keller**[3], **Sarah Comerford**[4], **Gail E. Tomlinson**[1,5], and **Yidong Chen**[1,6]

[1]Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

[2]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 5, ROC

[3]Pediatric Cancer Biology Program, Papé Family Pediatric Research Institute, Department of Pediatrics, Oregon Health and Science University, Portland, OR, USA

[4]Molecular Genetics, Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA

[5]Pediatric Department, School of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

[6]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

Abstract

Extracting maximal information from gene signature sets (GSSs) via microarray-based transcriptional profiling involves assigning function to up and down regulated genes. Here we present a novel sample scoring method called Signature-score (S-score) which can be used to quantify the expression pattern of tumor samples from previously identified gene signature sets. A simulation result demonstrated an improved accuracy and robustness by S-score method comparing with other scoring methods. By applying the S-score method to cholangiocarcinoma (CAC), an aggressive hepatic cancer that arises from bile ducts cells, we identified enriched oncogenic pathways in two large CAC data sets. Thirteen pathways were enriched in CAC compared with normal liver and bile duct. Moreover, using S-score, we were able to dissect correlations between CAC-associated oncogenic pathways and Gene Ontology function. Two major oncogenic clusters and associated functions were identified. Cluster 1, which included beta-catenin and Ras, showed a positive correlation with the cell cycle, while cluster 2, which included TGF-beta, cytokeratin 19 and EpCAM was inversely correlated with immune function. We also used S-score to identify pathways that are differentially expressed in CAC and hepatocellular carcinoma (HCC), the more common subtype of liver cancer. Our results demonstrate the utility and effectiveness of *S*-score in assigning functional roles to tumor-associated gene signature sets and in identifying potential therapeutic targets for specific liver cancer subtypes.

Keywords

gene signature set; pathway analysis; S-score method; tumor classification

Corresponding to: Yidong Chen. Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA. cheny8@uthscsa.edu.

Introduction

Cellular functions and biological processes are determined by cell-type-specific transcriptional programs. Thus, knowledge of the transcriptional derangements present in tumors can be used to predict altered function and identify pathways and processes that drive tumor growth. These altered functions and pathways can be targeted and exploited for therapeutic intervention. The inherent heterogeneity and complexity of tumors is refected by a significant degree of variability between gene signature sets (GSSs) obtained from similar tumor types. Gaining a broad-based view of pathways and processes associated with specific tumor type, thus requires further analysis such as group comparison or clustering of differentially expressed genes.

Different scoring methods have been developed to quantify the activities of GSSs associated with different disease states. Hiromichi and his colleagues used averaged *Z*-value of genes as a "deregulation index" to score GSS (1). Barbie *et al*. utilized Kolmogorov-Smirnov (KS) statistics to extend the Gene Set Enrichment Analysis (GSEA) (2) to project GSSs' expression to score single sample (3). Pearson correlation of comparing the directions of expression fold changes of member genes was also proposed for GSS scoring (4). Among these methods, *Z*-value and KS statistics are effective only when member genes have consistent directional changes in expression. However, these underlying assumptions do not necessarily hold true for the following reasons; (I) GSSs derived from genome-wide profiling experiments contain both up- and/or down-regulated components, which differ according to their cells and tissue of origin, and (II) the magnitude of gene expression alterations reflect magnitude of the response under any given experimental condition or disease state (e.g., benign *vs*. invasive tumors). In addition, many of these methods only consider the expression pattern (change in direction; up or down) and regard every gene with equal weight in the scoring scheme. In other words, the significance of each gene in the signature set is neglected. Recognizing these issues, we developed a unique scoring method to reconcile all of these factors within a GSS to generate a quantitative score.

To demonstrate the effectiveness of the method, we applied *S*-score to cholangiocarcinoma (CAC), the second most common primary liver cancer after hepatocellular carcinoma (HCC) (5). We applied our analysis to CAC as it is a clinically silent malignancy and as such, has an exceptionally dismal prognosis (5). Patients show a median survival rate of ~12 months with a 5-year survival rate of only 5–10% due to the presence of locally invasive or advanced metastatic disease at the time of diagnosis, precluding the possibility of resection. In cases where resection is possible, the high mortality rate reflects a high rate of post-surgical relapse and poor response to chemotherapy (6–8). While genetic alterations in KRAS, TP53, BRAF and EGFR have been reported for CAC (9–11), there has been a relative lack of systematic exploration. Thus, other alteration of important oncogenes and tumor suppressor genes were largely remain unclear.

To characterize individual CACs and identify oncogenic pathways and processes that drive biliary transformation, we used *S*-score, a novel signature scoring method adapted from our previously used scoring system (12) to gene expression profiles of CACs. *S*-score is different from previously used methods in that it concurrently evaluates both up- and down-regulated components of a GSS through a sign function and adjusts member genes' weights by differential expression significance, or *P*-value. As proof of principle and to demonstrate the accuracy and robustness of *S*-score, we conducted a thorough and comparative simulation under various conditions. In addition, the qualitative boundary of the *S*-score was established to provide a measure of the dynamic range of a GSS's activity (or the status of a oncogenic pathway). Finally, we applied *S*-score to two expression data sets of CAC containing 31 GSSs to identify putative oncogenic alterations, pathways and processes.

Using *S*-score, we identified several putative oncogenic alterations associated with CAC as well as differentially expressed pathways associated with CAC and HCC.

Merterials and methods

Signature score (S-score)

In order to project gene expression levels of a set of genes (a GSS) to a scalar score, a scoring method, adapted from our previous work (12), was developed to quantify each sample. We briefly describe the method as follows.

Suppose there are N genes in a GSS derived from a case-control gene expression profiling experiment. The case-control experiment contains M expression arrays, and $r_{j,i}$ represents the gene expression level of the jth gene from ith sample. Assuming there are W testing samples, and let $\mathbf{X}_l = \{x_{1,l}, \ldots, x_{N,l}\}$, where $x_{j,l}$ is the $\log_2$-transformed expression level of gene j in the testing sample l, $l = 1, \ldots, W$. To assess the activity of a gene set, we apply following operator,

$$s_l = \frac{1}{N}\sum_{j=1}^{N} Z_{j,l} \quad [1]$$

where $z_{j,l}$ is the z-score of a test sample, or

$$z_{j,l} = \frac{x_{j,l} - \mu_j}{\sigma_j^*} \quad [2]$$

where μ_j and σ_j^* are mean and standard deviation of gene j across all M case-control experiment samples. Similar approach has been utilized in [1] and [4] by replacing z-score with a Pearson correlation coefficient to case-control status in Eq. 1.

To incorporate the effect of differential expression significance and directionality, let ρ_j be the fold change of gene j between case and control, p_j be the P-value of Student t-test of gene j in the case-control experiment from which the GSS was derived. As defined before, N is the number of genes in the GSS. The *Signature Score* (*S*-score) of the testing sample l is defined as,

$$s_l = \frac{1}{K}\sum_{j=1}^{N} \mathrm{sign}\left(\rho_j\right) p_j^* z_{j,l} \quad [3]$$

where

$$p_j^* = \begin{cases} 3, & \text{if} -\log_{10} p_j > 4, \\ -\log_{10} p_j - 1, & \text{if} 1 \le -\log_{10} p_j \le 4, \\ 0, & \text{if} -\log_{10} p_j < 1, \end{cases} \quad [4]$$

and

$$K = \sum_{j=1}^{N} |p_j^*| \quad [5]$$

Equation [4] is set that if $p>0.1$, or $-\log_{10} p<1$, then $p^*=0$, and K [Eq.5] is the sum of all weight in Eq.[3]. p^* is designed to remove genes that has no test significance and weight more to differentially expressed genes, but limited to no more than $3\times$ weight when p is less

than 10^{-4}. Note in Eq. [3], if a test sample was profiled with the same microarray platform as that in case-control experiment, then,

$$\sigma_j^* = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_1^2}{n_1 + n_2 - 2}} \quad [6]$$

where σ_1 and σ_2 are the standard deviations of case and control, respectively, and n_1 and n_2 are the numbers of cases and controls in the experiment, respectively. However, if the test samples were profiled with a different microarray platform, μ_j and σ_j^* are mean and standard deviation of gene j across all W arrays of test samples, as an estimate of training samples.

Notice that S-score s_i, defined by Eq.[3], is positive when its standardized expression value z and the fold change of the corresponding gene in case/control samples have the same directional sign (positive or negative).

Qualitative boundary of S-score and no-call region

In concept, samples with positive S-scores that have expression patterns similar to case experiments can be classified as having an "active status" while those with negative scores can be classified as having an "inactive status". However, it is necessary to define a reliable prediction interval of active/inactive status in the whole dynamic range of S-score. In this study, we define the boundary of active/inactive statuses to be the 99% prediction interval under the null hypothesis that all signature genes are randomly distributed and not correlated to case/control in the conditional experiment. Let $z_{j,i}$ be normally distributed with $N[0, 1]$, and there are n genes and total of M microarrays balanced in active and inactive groups in the conditional experiment. The prediction interval is defined as,

$$L_{\alpha=0.01} = \mu_{s^*} \pm \Phi^{-1}_{N(0,1)}(0.01)\sigma_{s^*} = \mu_{s^*} \pm 2.58\sigma_{s^*} \quad [7]$$

where we assumed that the S-score of random gene expression s_i^* is approximately normally distributed [with standard cumulative distribution function $\Phi(\bullet)$] governed by the Large Number Theorem, μ_s and σ_s are mean and standard deviation of s_i^* which can be determined by simulation. For z^* to be $N[0, 1]$ in Eqs. [2–3], it can be easily shown that $\mu_s = 0$. The null hypothesis will be rejected and the status could be identified if the S-score exceeds or is below the $L_{0.01}$, or samples with an S-score between $L_{0.01}$ will not be classified in neither active nor inactive status (no-call group) to control the call-error below 1% due to random events. In Supplementary Materials section S3, we also provided a range of L values under various sample sizes, and a specific example for the determination of oncogene status with statistical confidential through our model [Eq.7]. Also, if the "no-call" region defined by $L_{0.01}$ covers entire dynamic region of a given GSS, the GSS will be classified as not involved in the biological system under study.

Generation of gene signature sets

A total of 31 GSSs were generated from data obtained from sources listed in Table S1 of the Supplementary Materials, The number of genes, data sources, and related studies were listed in Table S1 in the Supplementary Merterials. The cell cycle signature (CCS) that was used to define tumor cell proliferation was generated from a cell-cycle signature master set derived from genes preferentially expressed in actively cycling cells as well as genes that were repressed following growth inhibition (13,14). The wound healing signature (WHS) was derived from the gene sets response to serum stimulations (15). The KRT19 and EpCAM GSSs were generated from differentially expressed genes previously identified in

human or rat HCCs (16,17). c-Met and TGFβ signatures were generated from the expression profiles of wild-type and gene knockout mouse hepotocytes (18,19). The Shh (Sonic Hedgehog) signature was derived from an Shh-subtype of medulloblastoma (20). The RAF, MEK, Erbb2, and EGFR+EGF signatures were derived from genes found to be overexpressed in MCF-7 cells (21). The signatures of Src, E2F1, STAT3, p53, p63, Myc, ER, AKT, PI3K, Her2, TNF, IFNα, IFNβ, β-Catenin, EGFR, TGFβ, PR, ER, and Ras were derived from the study by Bild *et al.* (22).

Data set

Three datasets were used in this paper. The first data set which contains 94 embryonal rhabdomyosarcomas (eRMS) and 18 normal skeletal muscles, was download from caArray (https://array.nci.nih.gov/caarray/project/detailsaction?project.experiment.publicIdentifier=trich-00099) that we used in (12). Data set GSE26566 and GSE15765 were downloaded from GEO database. GSE26566 contains 104 cholangiocarcinoma specimens (CAC), 59 normal liver tissues (NL), and 6 normal intrahepatic bile ducts (IBD). GSE15765 contains 70 hepatocellular carcinoma (HCC), 13 cholangiocarcinoma (CAC), and 7 samples of mixed type of combined HCC and CAC (CHC).

Results

Performance and robustness of S-score

To demonstrate the performance of *S*-score, the eRMS data, discussed in (12), were used to compare the capability of *S*-score against the other three scoring methods, *ES*-score (3), averaged *Z*-score (1), and Pearson correlation (4). The result showed that (see Figure S1), among 4 methods, *S*-score and correlation-based method have the capacity to integrate up- and down-regulated gene sets. The two methods can accurately project expression pattern to a signature score, with the suppression of the noisy effect of unrepresentative genes in the GSS, if they exist (Details in Supplementary Materials section *S1*).

To further estimate the robustness of the scoring methods, we performed simulations where we added random noises into one *positive sample* as marked in Figure S1A. The simulation was performed on one of GSSs, p53off, derived by comparing cancer samples with p53 loss of function *vs.* samples with normal p53 function. For *Z*-score method, we separate p53 off GSS into two sets: p53off-up and p53off-down where up- and down-regulated genes were grouped, respectively. The mean of added noises was set to zero and the strengths (standard deviation) were increased from 0.1 to 2 times the standard deviation of each gene. Each condition was simulated 1,000 times and then calculations were made for the mean shift and standard deviation of each score. The result, shown in Figure 1, revealed that *S*-score and *Z*-score have the smallest score shift (all close to zero), regardless of the standard deviation of added noise. The mean shift of correlation is similar to *ES*-score in p53off-up and p53off-down that increase with the standard deviation of noise. Among these methods, *ES*-score with p53off-up gene set has the largest mean shift in the simulations, indicating worst robustness under noisy conditions (Figure 1A). The standard deviations of all four test scores were varied at a similar range (Figure 1B), with those of *ES*-score with p53off-up genes and correlation slightly lower than other methods. However, these two methods also have the largest mean-shift under noisy conditions.

The robustness of *S*-score was further demonstrated in Supplementary Materials Section S2 when a *negative sample* was also applied for the simulation. Similarly, *S*-score have a small mean shift and a similar range of standard deviation, as shown in Figure S2. Combining these two observations from simulations, *S*-score and *Z*-score have the best robustness

(mean shift close to zero) against noises in the test samples, while *S*-score has the ability of suppressing non-performing genes in a GSS.

Identification of oncogenic pathways of cholangiocarcinoma

Cholangiocarcinoma (CAC) is one of malignent cancers in the world, however, only mutations of gene Kras, p53 and Raf have been reproted. Other driving oncogentic pathways are still unclear. Here we ultized *S*-score method to investgate the activities of oncogenic pathways of CAC through gene expression profiles. To identify the driving oncogene, oncogenic signatures of known oncogenes and the actived pathway have been reported in the HCC, were collected. Total 31 oncogenic signatures was applied to a gene expression data set of CAC, GSE26566, based on the *S*-score to estimate the activities of the pathways. The data set contains 104 CAC specimens, 59 normal liver tissues (NL), and 6 normal intrahepatic bile ducts (IBD). Three gene signature sets (GSSs), p63, AKT, and PR, were excluded due to none or only one sample passed the qualitative boundary $L_{0.01}$. The c-Met signature was also excluded because most of the up or down-regulated genes in the GSS were highly expressed in NL samples that lost the trend of expression direction in the original case-control study (Figure S4 in Supplementary Materials). *S*-scores for each sample and each GSS were generated and then a Student *t*-test was used to identify the statistical significance of differential oncogenic pathways. Using *P*-values of *t*-test smaller than 0.05 as the criteria, 13 of 27 signatures were shown the differential activities between CAC and the two normal tissues (Figure 2 and S5 in the Supplementary Materials). Twelve out of 13 GSSs showed higher values of *S*-score in CAC. Only Src signature had lower value of *S*-score in CAC. On average, 21 CAC samples (20%) have identified as “on” status of signatures, which labeled as cyan in the Figure 2. KRT19 have the highest identified percentage: 73 of 104 CAC samples were identified as “on” status of KRT19 with an *S*-score greater than $L_{0.01}$ (0.44). The result was consistent with previous reports that KRT19 is highly expressed in CAC (23–25).

To investigate the co-regulation of the oncogenic signatures and the subtypes of CAC, a two-way hierarchical clustering was performed. The result identified two major clusters of signatures’ co-activity. The signature cluster A, which labeled with blue color in Figure 3A, manifested the correlation of β-Catenin, Ras, CCS and wound signatures. The cluster B which labeled with red color exhibited the co-expression of TGFβ, KRT19, and EpCAM. The values of Pearson correlation between two GSSs were further substantiated the clustering results (Figure 3B). The data also showed that IFNα and Src signatures were active in an independent manner without correlation with other signatures. On the other hand, 6 clusters of CAC samples were identified by using hierarchical clustering (Figure 3). The cluster 1 has the strongest activities (except of IFNα and Src signatures) among all the clusters; Cluster 2 has less activity of signature cluster A; Cluster 3 has less expression of p53 and MEK; Cluster 4 has high activities of signature cluster B with different combination of other signature activities; Cluster 5 has moderate activities of signature cluster B but had low activities in most of other signatures; and Cluster 6 shows low activities of all signatures (except INFα and Src). Clearly, our result demonstrated the heterogenetic characteristic of oncogenic background in CAC patients.

The oncogene associated gene ontology items of cholangiocarcinoma

To understand the dysregulated function of CACs, we also investigated GSSs with Gene Ontology (GO) terms. Using $\log_2$-fold change >0.4 and *P*-value <0.01 in both CAC *vs*. NL and CAC *vs*. IDB comparisons as the selecting criteria, a total of 68 GO terms, containing 39, 12, and 17 terms in biological process (BP), molecular function (MF), and cellular components (CC), respectively, were shown significant differentially activities by *S*-score in CAC (Table S2 in Supplementary Materials). In the CAC associated GO items, 54 of 68

CAC associated GO items were up-regulated (32, 10, and 12 in BP, MF, and CC, respectively) and 14 were down-regulated (5, 2, and 2 in BP, MF, and CC, respectively). Several mitotic correlated terms, such as mitotic spindle organization, mitotic prometaphase, and M phase of mitotic cell cycle, were up-regulated in CACs. Also, RNA transcription related terms and telomere maintenance terms were up-regulated as well. Those results indicated CACs have strong cell cycle and RNA transcription activities comparing with normal NL or IDB tissues. We also observed down-regulation of GO terms of triglyceride homeostasis, complement activation, acute-phase response, and lipid transporter, indicating the potential loss of these functions in CAC cells.

To investigate the relationship between the CACs associated oncogene pathways and functions, Pearson correlation was applied to the *S*-score values of the 12 oncogenes and 68 GO terms. A total 45 terms met the criteria of correlation coefficients >0.7. The cell cycle and mitotic related terms were highly correlated with the cluster A (Table 1). Also other terms of cell functions, such as mRNA metabolic, translation, and telomere maintenance via recombination, possessed positive correlation with signature cluster A. This association suggested the activities of β-Catenin and Ras, the oncogenes of the Cluster A, were related to activities of cell dividing, transcription, and translation in CACs. Similarly, due to the negative correlation of the Cluster B to two immune related terms, acute-phase response and complement activation, and lipoprotein related terms, such as lipid transporter activity and triglyceride homeostasis, suggested involvement of Cluster B in the suppression of the immune response and lipid transporter functions (Table 2).

Comparsion of cholangiocarcinoma and hepatocellular carcinoma

The hepatocellular carcinoma (HCC) with cholangiocarcinoma signature has been reported with worse prognosis (26). However, the differences between these two cancers in the oncogenic pathway are still unknown. A data set GSE15765 that contains 70 hepatocellular carcinoma (HCC), 13 cholangiocarcinoma (CAC), and 7 samples of mixed type of combined HCC and CAC (CHC) were used to investigate the differences with our oncogenic GSSs. By performing *t*-test between HCC and CAC samples, total 12 oncogenic GSSs were different with statistical significance $P<0.05$. Nine of 12 signatures have higher activities in CAC (Figure 4A) and 3 oncogenic GSSs, AKT, PI3K, and E2F1, showed greater levels in HCC (Figure 4B). Among the 9 highly expressed GSSs in CAC, 6 were overlapped with the analysis of GSE26566: MEK, Ras, k-ras addiction, KRT19, EpCAM, and TGFβ. The GSSs of Her and Erbb2, which were signature of the same gene, *ERBB2*, but derived from different experiments, were both shown highly expressed in CAC.

In addition to HCC, the mixed tumor type CHC showed mixed oncogenic GSSs *S*-score range (activities) in all 12 GSSs, as we expected. To summarize our observation in Figure 4, HCC and CAC have distinct activities of oncogenic pathways, and they can be efficiently utilized for tumor classification, instead of individual genes. However, the topic is beyond the scope of this paper.

Discussion

We proposed a novel tumor scoring system, *S*-score, to quantify oncogenic GSSs in tumor comparison. By considering the directions of fold change and P-values of GSSs, *S*-score provides an effective scoring method for each tumor with accuracy and robustness, as demonstrated in the Result Section. Not only for quantitation, by estimating the confidential boundary $L_{0.01}$, the qualitative status of GSSs of each sample can also be determined. The boundary $L_{0.01}$ was also utilized to filter out the signatures of which distribution of *S*-scores are relatively small. For example, the results of the signatures p63, AKT, and PR, were

excluded in the analysis of GSE26566 after filtering by $L_{0.01}$ to avoid the situation of small variance of *S*-score, causing instable *t*-test significance.

Through *S*-score, we expect further insight into gene signature set analysis of a tumor to understand the regulation and function of pathway of cancer biology. Using expression profiles of CACs, 12 oncogenic GSSs were shown with high activities in CAC comparing with two normal tissues (Figure 2), consistent with the observations reported earlier in literatures. Kras and p53 have been reported with high frequency of mutation in the CAC patients (9,10). Mutation of *SMAD4*, which was a downstream gene of TGFβ pathway, also has been identified in a proportion of CAC patients (27). Immunohistochemical analysis of β-Catenin in a previous study showed positive staining in cytoplasm and/or nucleus in 58.3% of 24 CAC samples (28). KRT19, EpCAM were highly expressed in biliary epithelial cells, and have been report as the markers of hepatic progenitor cells with β-Catenin (16,17). Our data indicated that KRT19 and EpCAM were coexpressed with TGFβ in CAC. Although no mutations or alteration of expression of MEK in CAC have been reported earlier, but a phase II clinical trial have been reported with 12% response rate of selumetinib, a MEK inhibitor, for CAC patients (29).

Our analysis also showed that cell-cycle signature (CCS) and wound healing signature were highly correlated with GSSs of Myc and β-Catenin in CACs (Figure 3). The high scores of CCS and wound signatures indicated high proliferation and invasion in most of CAC samples, perhaps Myc and/or β-Catenin play important roles in the tumor progression of CAC.

Two Ras related GSSs were showed associated with CACs (Figure 2). However, the correlation of the GSSs of Ras and K-ras addiction was low (Figure 3). The Ras GSS was derived from gene affected by Hras overexpression. The Kras addiction was derived from the K-ras dependent cell lines, which mean K-ras played a dominant role of the proliferation signal in the cell lines. While we don't expect two GSSs have the similar *S*-scores for all CACs, we assume other oncogenes, perhaps not Kras, may play the more significant roles even in the Kras mutated CAC tumors. By analyzing all oncogenic GSSs, our analysis may provide additional clue to other oncogenes that CACs may addict to.

The cluster analysis in the study identified 6 subclasses of CAC patients with different combinations of oncogene activities. This information can provide a useful molecular evidence for predicting prognosis and making treatment decisions, especially for target drugs. For example, the patients with higher MEK activity could have high probability to response to MEK inhibitor (29). Additional treatment strategies can be designed to target other oncogenic pathways with the potential to increase the drug response rate toward the goal of personalized medicine. For example, in the CACs cluster 6, the low activities of GSSs not only observed in cluster B (TGFβ, KRT19, and EpCAM) but also in all other GSSs except Src. We assumed that there were other oncogenic pathways (GSSs) not collected in the study. As we can see from Figure 3A, more GSSs are needed to uncover the known oncogenes in patients of clusters 2–4.

The correlation between GSSs and GO terms were also investigated in this study. The cell cycle related GO terms were associated with GSSs of the Cluster A. The result further confirmed that two oncogenes, Myc and β-Catenin, might play an important role in the cell cycling of CAC cancer cells. In addition, the two oncogenes were also associated with the activities of other essential function of cells, such as mRNA metabolic, translation, and telomere maintenance via recombination, suggesting the causal relationships might exist between the two oncogenes and the functional consequences. Our analysis also indicated the activities of the Cluster B (EpCAM, KRT19, and TGFβ) was negatively correlated to two

immune functions, acute-phase response and complement activation. Since the chronic inflammation has been identified as an important factor of carcinogenesis (30,31), not only these three genes are used as markers for diagnosis of CACs, but also are involved in the carcinogenesis of CAC due to their ability to suppress immune functions.

Comparing to HCC at oncogenic GSS level, CACs have higher activities of MEK and Her2 while HCCs show high activities of AKT, PI3K and E2F1. Timothy H. *et al*. reported that PI3K/AKT signaling but not MEK blocks E2F1-induced apoptosis and switch E2F1 to proliferative function (32). Aberrant Her2 expression in the CAC patients with poor prognosis has been described by Jesper A. *et al*. (33). Our results, combined with literature reports, clearly suggested the proliferation signaling between HCC and CAC is different. Moreover, HCC patients with high activities of EpCAM, KRT19, and TGFβ have been reported with poor prognosis. These findings supported the clinical observations that CAC is malignant with relatively short survival rate.

The progression of cancer can be reasoned from different kinds of alteration in the genome, such like mutations, copy number variations, and epigenetic modifications. The consequences of these alterations will induce changes of gene expression and then to the protein expression level in their pathways. Utilizing our *S*-score method that estimates the expression pattern of the oncogenic GSSs derived from *in vivo* or *in vitro* experiments, the putative oncogenic pathways could be uncovered as we demonstrated in the two CAC data sets. Some of the oncogenic pathways we identified were well known and some are yet characterized in CAC. Those data demonstrated that *S*-score is a rational strategy to explore novel oncogenic pathway for microarray data. Not only limit to the 31 oncogenic GSSs, our method can incorporate with other gene sets, such like the gene sets collected in MSigDB, to provide a systematic examining of the activities of pathways studied in previous literatures. The *S*-score values of gene sets can also incorporate with other clinical information, such like survival information and stage status, with different statistical methods for different proposes of studies in cancer. The flexibility of *S*-scare will be useful to uncover novel mechanism and pathway in cancer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Disclosure: This work was supported by the NIH/NCI cancer center grant (P30 CA054174-17), NIH/NCRR CTSA grant (1UL1RR025767) to YC, Cancer Prevention and Research Institute of Texas grant (CPRIT RP101195-C04) to GET, SC, and YC. TH is supported by the Greehey Children Cancer Research Institute (GCCRI) intramural research fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Ebi H, Tomida S, Takeuchi T, et al. Relationship of deregulated signaling converging onto mTOR with prognosis and classification of lung adenocarcinoma shown by two independent in silico analyses. Cancer Res. 2009; 69:4027–4035. [PubMed: 19383916]
2. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–15550. [PubMed: 16199517]
3. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009; 462:108–112. [PubMed: 19847166]
4. Gibbons DL, Lin W, Creighton CJ, et al. Expression signatures of metastatic capacity in a genetic mouse model of lung adenocarcinoma. PLoS One. 2009; 4:e5401. [PubMed: 19404390]

5. Lazaridis KN, Gores GJ. Cholangiocarcinoma. Gastroenterology. 2005; 128:1655–1667. [PubMed: 15887157]
6. Eckel F, Schmid RM. Chemotherapy in advanced biliary tract carcinoma: a pooled analysis of clinical trials. Br J Cancer. 2007; 96:896–902. [PubMed: 17325704]
7. Yusoff AR, Razak MM, Yoong BK, et al. Survival analysis of cholangiocarcinoma: a 10-year experience in Malaysia. World J Gastroenterol. 2012; 18:458–465. [PubMed: 22346252]
8. Shaib Y, El-Serag HB. The epidemiology of cholangiocarcinoma. Semin Liver Dis. 2004; 24:115–125. [PubMed: 15192785]
9. Ohashi K, Nakajima Y, Kanehiro H, et al. Ki-ras mutations and p53 protein expressions in intrahepatic cholangiocarcinomas: relation to gross tumor morphology. Gastroenterology. 1995; 109:1612–1617. [PubMed: 7557145]
10. Furubo S, Harada K, Shimonishi T, et al. Protein expression and genetic alterations of p53 and ras in intrahepatic cholangiocarcinoma. Histopathology. 1999; 35:230–240. [PubMed: 10469215]
11. Tannapfel A, Sommerer F, Benicke M, et al. Mutations of the BRAF gene in cholangiocarcinoma but not in hepatocellular carcinoma. Gut. 2003; 52:706–712. [PubMed: 12692057]
12. Rubin BP, Nishijo K, Chen HI, et al. Evidence for an unanticipated relationship between undifferentiated pleomorphic sarcoma and embryonal rhabdomyosarcoma. Cancer Cell. 2011; 19:177–191. [PubMed: 21316601]
13. Mizuno H, Nakanishi Y, Ishii N, et al. A signature-based method for indexing cell cycle phase distribution from microarray profiles. BMC Genomics. 2009; 10:137. [PubMed: 19331659]
14. Cam H, Balciunaite E, Blais A, et al. A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell. 2004; 16:399–411. [PubMed: 15525513]
15. Chang HY, Sneddon JB, Alizadeh AA, et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol. 2004; 2:E7. [PubMed: 14737219]
16. Andersen JB, Loi R, Perra A, et al. Progenitor-derived hepatocellular carcinoma model in the rat. Hepatology. 2010; 51:1401–1409. [PubMed: 20054870]
17. Yamashita T, Forgues M, Wang W, et al. EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. Cancer Res. 2008; 68:1451–1461. [PubMed: 18316609]
18. Lamb JR, Zhang C, Xie T, et al. Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS One. 2011; 6:e20090. [PubMed: 21750698]
19. Coulouarn C, Factor VM, Thorgeirsson SS. Transforming growth factor-beta gene expression signature in mouse hepatocytes predicts clinical outcome in human cancer. Hepatology. 2008; 47:2059–2067. [PubMed: 18506891]
20. Kool M, Koster J, Bunt J, et al. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One. 2008; 3:e3088. [PubMed: 18769486]
21. Creighton CJ, Hilger AM, Murthy S, et al. Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an in vivo molecular phenotype of estrogen receptor alpha-negative human breast tumors. Cancer Res. 2006; 66:3903–3911. [PubMed: 16585219]
22. Gatza ML, Lucas JE, Barry WT, et al. A pathway-based classification of human breast cancer. Proc Natl Acad Sci U S A. 2010; 107:6994–6999. [PubMed: 20335537]
23. Jain R, Fischer S, Serra S, et al. The use of Cytokeratin 19 (CK19) immunohistochemistry in lesions of the pancreas, gastrointestinal tract, and liver. Appl Immunohistochem Mol Morphol. 2010; 18:9–15. [PubMed: 19956064]
24. Lie-A-Ling M, Bakker CT, Deurholt T, et al. Selection of tumour specific promoters for adenoviral gene therapy of cholangiocarcinoma. J Hepatol. 2006; 44:126–133. [PubMed: 16168519]
25. Balaton AJ, Nehama-Sibony M, Gotheil C, et al. Distinction between hepatocellular carcinoma, cholangiocarcinoma, and metastatic carcinoma based on immunohistochemical staining for carcinoembryonic antigen and for cytokeratin 19 on paraffin sections. J Pathol. 1988; 156:305–310. [PubMed: 2465399]

26. Woo HG, Lee JH, Yoon JH, et al. Identification of a cholangiocarcinoma-like gene expression trait in hepatocellular carcinoma. Cancer Res. 2010; 70:3034–3041. [PubMed: 20395200]
27. Argani P, Shaukat A, Kaushal M, et al. Differing rates of loss of DPC4 expression and of p53 overexpression among carcinomas of the proximal and distal bile ducts. Cancer. 2001; 91:1332–1341. [PubMed: 11283934]
28. Tokumoto N, Ikeda S, Ishizaki Y, et al. Immunohistochemical and mutational analyses of Wnt signaling components and target genes in intrahepatic cholangiocarcinomas. Int J Oncol. 2005; 27:973–980. [PubMed: 16142313]
29. Bekaii-Saab T, Phelps MA, Li X, et al. Multi-institutional phase II study of selumetinib in patients with metastatic biliary cancers. J Clin Oncol. 2011; 29:2357–2363. [PubMed: 21519026]
30. Palmer WC, Patel T. Are common factors involved in the pathogenesis of primary liver cancers? A meta-analysis of risk factors for intrahepatic cholangiocarcinoma. J Hepatol. 2012; 57:69–76. [PubMed: 22420979]
31. Fava G. Molecular mechanisms of cholangiocarcinoma. World J Gastrointest Pathophysiol. 2010; 1:12–22. [PubMed: 21607138]
32. Hallstrom TC, Mori S, Nevins JR. An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. Cancer Cell. 2008; 13:11–22. [PubMed: 18167336]
33. Andersen JB, Spee B, Blechacz BR, et al. Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. Gastroenterology. 2012; 142:1021–1031. e15. [PubMed: 22178589]

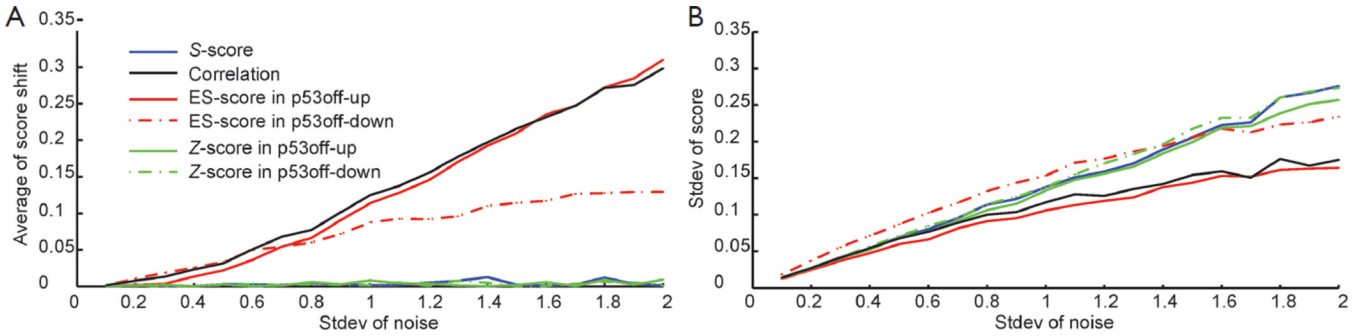

Figure 1.

The plots of scoring variation under noise perturbation. The positive score ΔES ΔES samples, as the red arrows marked in Figure S1, were applied for the simulations of robustness under noise disturbance. (A) The mean shifts, and (B) standard deviations of the sample with positive score (p53 active)

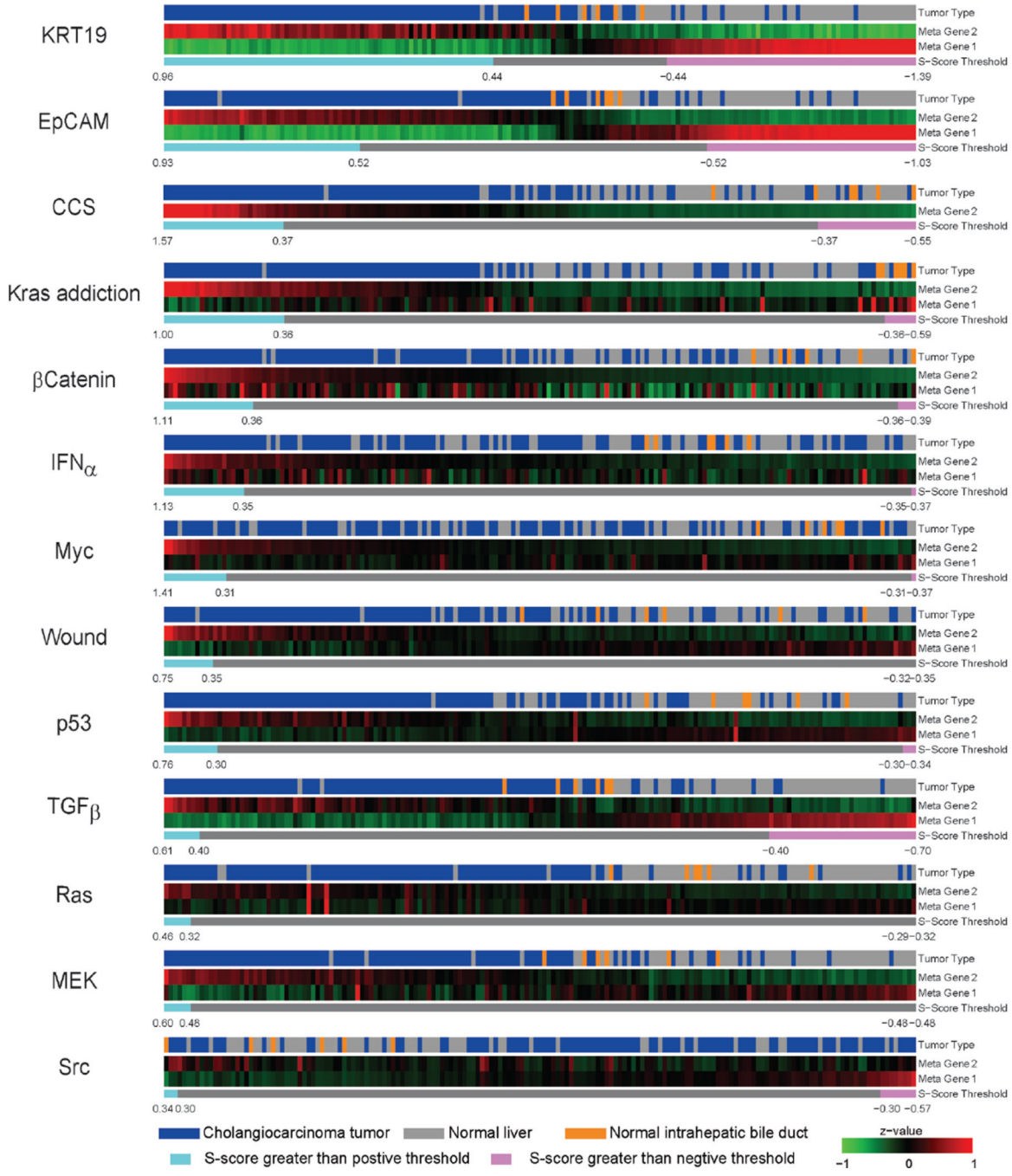

Figure 2.

CAC associated oncogenic signatures. Thirteen differential expressed signatures have been identified. Twelve of them were up-regulated in CAC and only Src was down regulated. The meta genes 1 and 2 are the averaged expression level of down and up regulated gene in the signature, respectively

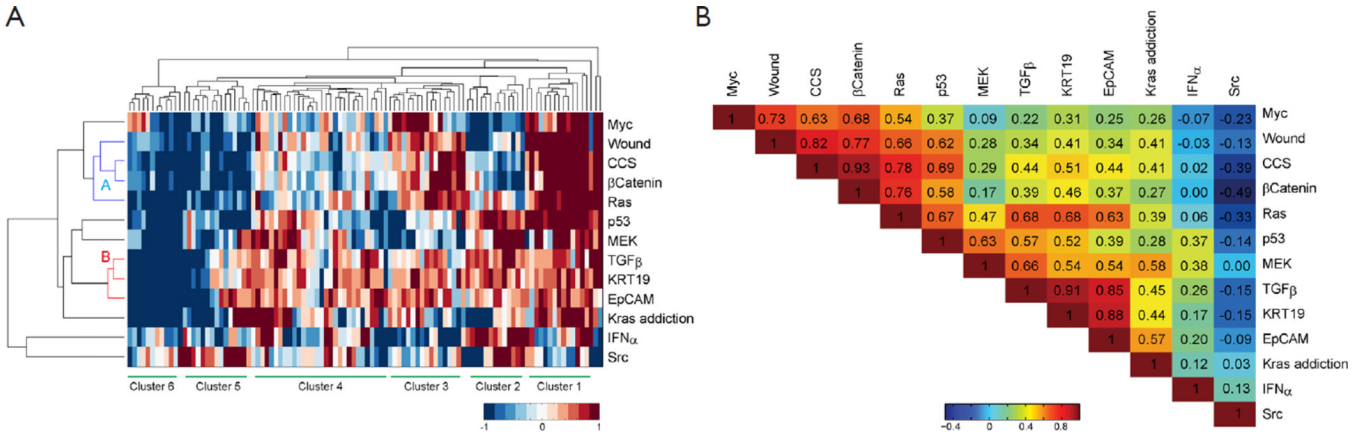

Figure 3.

The activity profile of the oncogenic pathway in cholangiocarcinoma. (A) The clustergram of the signature scores. The score reflected the activities of the oncogenic pathway in each CAC patients. Two clusters of oncogenic pathway were reveal with coregulation and also 6 groups patients with different pattern of pathway label were identified, and (B) The correlation heatmap. The heat map depicting the correlation coefficient of pathway coregulation

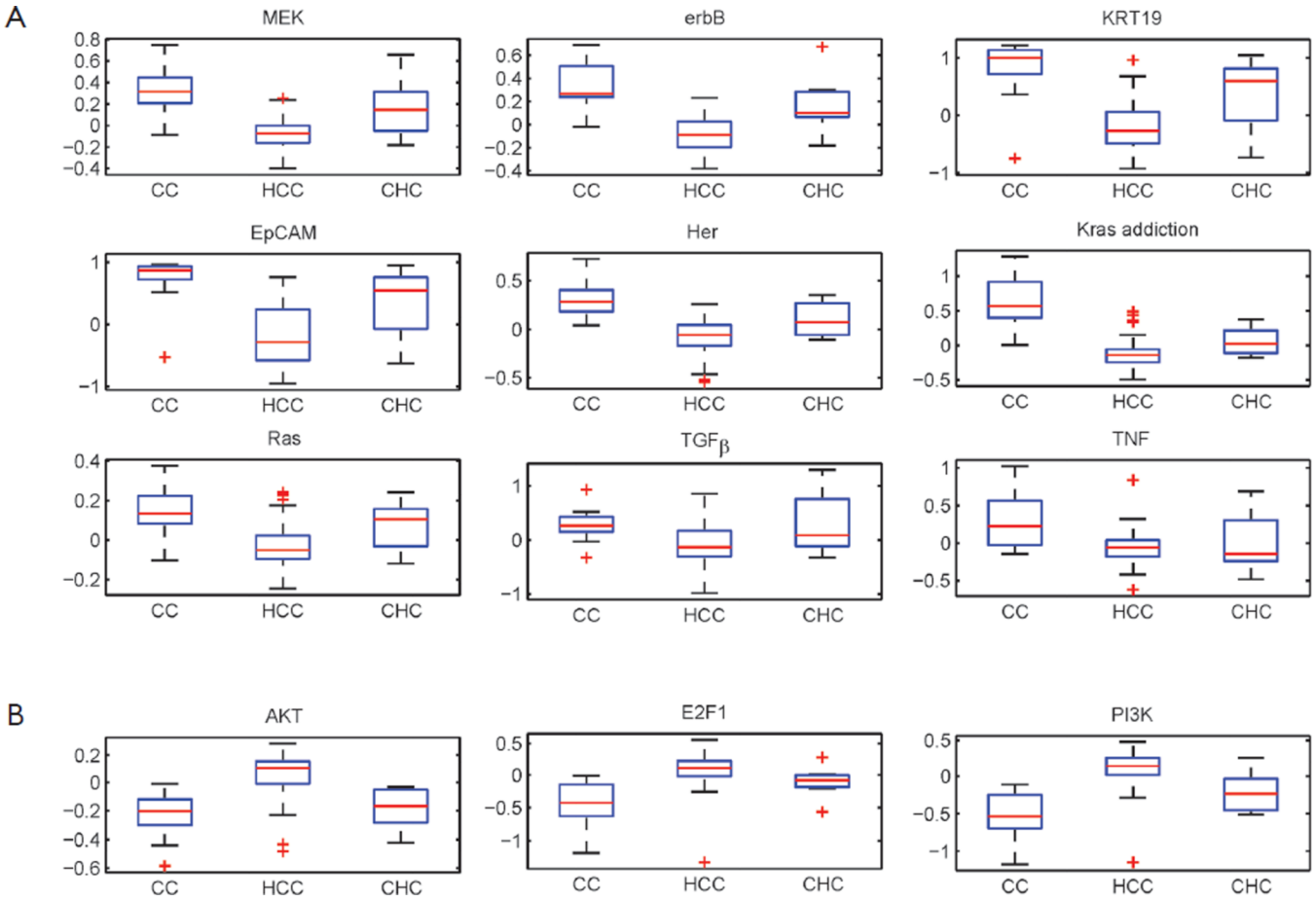

Figure 4.

The boxplot of the 12 signatures. Total 12 signatures showed differential activities between cholangiocarcinoma (CAC) and hepatocellular carcinoma (HCC). (A) 9 signatures were up-regulated in CAC, and (B) 3 showed down regulation. The *P*-values of *t*-test of the 12 signatures were all smaller than 0.01. CHC is the mixed type of combined HCC and CAC

Table 1

Gene ontology terms correlated with cluster A (Ras, b-catenin, CCS, Wound healing)

Type	CAC vs. NL ΔS	CAC *vs.* IBD ΔS	Correlation	Gene set
BP	0.49	0.55	0.95	Mitotic cell cycle
	0.41	0.45	0.95	Cell division
	0.56	0.66	0.95	M phase of mitotic cell cycle
	0.57	0.66	0.95	Mitotic prometaphase
	0.70	0.75	0.89	Mitotic spindle organization
	0.47	0.54	0.88	Cell cycle checkpoint
	0.67	0.75	0.88	DNA strand elongation involved in DNA replication
	0.49	0.64	0.87	mitotic sister chromatid segregation
	0.48	0.66	0.86	CenH3-containing nucleosome assembly at centromere
	0.49	0.66	0.86	Telomere maintenance via recombination
	0.52	0.56	0.86	Regulation of transcription involved in G1/S phase of mitotic cell cycle
	0.43	0.53	0.85	S phase of mitotic cell cycle
	0.50	0.67	0.85	Telomere maintenance via semi-conservative replication
	0.43	0.51	0.83	Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
	0.42	0.59	0.82	Nucleotide-excision repair, DNA gap filling
	0.46	0.54	0.81	M/G1 transition of mitotic cell cycle
	0.42	0.50	0.80	Regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
	0.47	0.67	0.80	DNA-dependent DNA replication initiation
	0.41	0.49	0.79	Double-strand break repair via homologous recombination
	0.49	0.49	0.78	Mitotic cell cycle spindle assembly checkpoint
	0.65	0.78	0.77	Chromosome organization
	0.44	0.47	0.73	Negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
	0.53	0.41	0.71	RNA metabolic process
MF	0.48	0.56	0.76	ATP-dependent DNA helicase activity
	0.41	0.53	0.75	Nuclease activity
	0.47	0.67	0.71	DNA helicase activity
CC	0.45	0.56	0.93	Condensed chromosome kinetochore
	0.41	0.53	0.89	Kinetochore
	0.44	0.40	0.89	Apindle pole
	0.42	0.57	0.87	Condensed chromosome
	0.52	0.53	0.85	Spindle microtubule
	0.46	0.42	0.77	Ribonucleoprotein complex
	0.55	0.47	0.72	U12-type spliceosomal complex

Note: CAC, cholangiocarcinoma; NL, normal liver; IDB, intrahepatic bile duct; BP, Biological process; MF, Molecular function; CC, Cellular component; ΔS difference of S-score

Table 2

Gene ontology items correlated with cluster B and Myc

Type	CAC *vs.* NL ΔS	CAC *vs.* IBD ΔS	Correlation	Gene set
Cluster B (TGFβ, KRT19, EpCAM)				
BP	−0.60	−0.50	−0.83	acute-phase response
	−0.69	−0.63	−0.86	triglyceride metabolic process
	−0.95	−0.58	−0.88	blood coagulation, intrinsic pathway
	−1.03	−0.63	−0.92	complement activation
	−1.09	−0.70	−0.93	triglyceride homeostasis
MF	−0.59	−0.51	−0.75	fatty acid binding
	−0.89	−0.57	−0.89	lipid transporter activity
CC	−0.92	−0.57	−0.86	very-low-density lipoprotein particle
	−0.99	−0.68	−0.88	high-density lipoprotein particle
Myc				
BP	0.43	0.41	0.72	translation
CC	0.63	0.48	0.77	small nuclear ribonucleoprotein complex
	0.46	0.42	0.70	ribonucleoprotein complex

Note: CAC, cholangiocarcinoma; NL, normal liver; IDB, intrahepatic bile duct; BP, Biological process; MF, Molecular function; CC, Cellular component; ΔS difference of S-score