INTERNATIONAL JOURNAL OF
MEDICAL BIOCHEMISTRY

DOI: 10.14744/ijmb.2022.18199
Int J Med Biochem 2022;5(3):137-151



Research Article

Identification of key genes and pathways for cholangiocarcinoma using an integrated bioinformatics analysis

Asli Kutlu[1], Merve Arda[2], Evren Atak[3], Engin Ulukaya[4]

[1]Department of Bioinformatics & Genetics, Istinye University Faculty of Engineering and Natural Science, Istanbul, Türkiye
[2]Department of Cancer Biology and Pharmacology, Istinye University, Health Science Institute, Istanbul, Türkiye
[3]Department of Bioinformatics and System Biology, Gebze Technical University, Institute of Natural and Applied Sciences, Kocaeli, Türkiye
[4]Department of Clinical Biochemistry, Istinye University Faculty of Medicine, Istanbul, Türkiye

Abstract

Objectives: The scope of this study was to identify potential genes as a promising biomarker in diagnosing cholangiocarcinoma (CCA) or differentiating the subtypes of CCA. In this study, we used Gene Expression Omnibus (GEO)-NCBI data sets as promising open sources to perform integrative analysis.

Methods: The gene expression data sets of intrahepatic CCA (iCCA) and extrahepatic CCA (eCCA) were retrieved from GEO, and the statistical analysis of GSE45001 (iCCA), GSE76311 (iCCA), and GSE132305 (eCCA) was performed to identify significantly expressed genes. The association of listed genes with CCA was checked via text-mining approaches. For CCA, the details were provided by discussing its relations with our results. Then, the pathway analysis was performed to identify common pathways both in iCCA and eCCA.

Results: The pathway analysis reveals that although there are common pathways between iCCA and eCCA, the associated genes within these pathways are different from one another. According to the results of upregulated gene sets, integrin cell surface interaction (R-HSA-216083), MET activates PTK2 signaling (R-HSA-8874081), degradation of the extracellular matrix (ECM) (R-HSA-1474228), nonintegrin membrane–ECM interaction (R-HSA-3000171), and assembly of collagen fibrils and other multimeric structures (R-HSA-2022090) are found as common pathways among these data sets, yet there is no reported common pathway within downregulated gene sets. A detailed study of common pathway analysis shows that *COL1A1* and *COL1A2* genes, whose associations with CCA have not been reported, seem promising to differentiate iCCA from eCCA. The pathway analysis also reveals that although there are common pathways between iCCA and eCCA, the associated genes within these pathways are different from one another.

Conclusion: Focusing on pathways rather than genes is more promising for revealing the potential biomarkers together with providing a deeper understanding by highlighting significant pathways.

Keywords: *COL1A1*, *COL1A2, eCCA*, gene expression study, GEO data sets, iCCA

Cholangiocarcinoma (CCA) is a rare malignant cancer arising from extrahepatic and intrahepatic biliary epithelial cells, and it accounts for 10%-20% and 3% of primary and gastrointestinal cancer types, respectively, around the world [1]. Its prevalence is reported as 0.5-1.2 out of 100 000 as being higher gender incidence in men compared with women. There are three groups in CCA according to anatomical localization: (1) intrahepatic, (2) perihilar, and (3) distal extrahepatic [2]. CCA has a low 5-year survival rate upon surgery and chemotherapy treatments [3, 4]. Although great efforts are made during its routine diagnosis, only 1 out of 3 CCA patients are diagnosed in the early stage, which is low compared with that of other cancers.

Address for correspondence: Asli Kutlu, MD. Department of Bioinformatics & Genetics, Istinye University Faculty of Engineering and Natural Science, Istanbul, Türkiye
Phone: +90 534 772 33 87 **E-mail:** asli.kutlu@istinye.edu.tr **ORCID:** 0000-0002-9169-388X
Submitted Date: June 25, 2022 **Accepted Date:** August 16, 2022 **Available Online Date:** September 15 2022
OPEN ACCESS This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

In recent decades, there have been great contributions from advanced molecular techniques to the field of cancer diagnosis and treatment by enabling patient-specific molecular profiling and integrating *in vivo* and *in vitro* finding with clinics [5, 6]. The diagnosis stage of cancer is a crucial parameter for treatment response in which tumor markers are used [7]. In addition to early diagnosis, tumor markers are also used for screening, staging, or disease monitoring. The accuracy and efficiency of tumor markers are crucial parameters to widen their usage because they define the risk of overdiagnosing. The World Health Organization defines a biomarker as "a process, outcome or incidence of disease that can be measured in any substance, structure or body or its products and which can affect or predict the functioning of the body" [8]. Similarly, tumor markers are defined as biomarkers whose increased expression level is in close association with cancer.

As cancer is considered a complex disease in which its contributors are varied from one person to another, the discovery of novel biomarkers for each type of cancer is continuously demanded. The lessons coming from personal medicine in cancer taught us that there is patient-specific variance in treatment, leading to a shift in biomarker discovery studies rather than proposing specific gene, miRNA, and protein to propose a pathway-specific biomarker. It emphasizes the importance of pathway specificity for either selected cancer type or subtype by decreasing the errors during the screening stage of different populations to test the specificity of selected biomarker(s). This approach is in line with what we have learned from the personalized medicine approach that states the accumulation of mutation(s) in pathways rather than specific gene(s) as being an actual driving force for cancer. Until now, several CCA-associated biomarkers, such as cysteine dioxide type 1 (*CDO1*) [9], secreted curvy protein 1 (*SFRP1*) [10], zinc finger and SCAN domain protein 1 (*ZSCAN18*) [11], and cool/threonine-protein kinase 1 (*DCLK1*) [12], are reported with limited usage due to lack of specificity and accuracy. Because these genes do not exist in the same or associated pathways, it would decrease their specificity and accuracy parameters.

In this work, we aim to discover novel biomarkers by using GEO data sets of iCCA (GSE45001 and GSE76311) and eCCA (GSE132305) patients via a pathway-specific approach. Here, we report five common pathways between iCCA and eCCA. Through the integration of pathway analysis with a statistical approach, we detect *COL1A1* and *COL1A2* genes as promising biomarkers to differentiate iCCA from eCCA, and their association with CCA is reported in our work. This study also demonstrates the power of a pathway-based approach to discover the potential biomarkers that could be used to differentiate subtypes of CCA.

Materials and Methods

Retrieving data sets and processing with R-language

We used GEO data sets of GSE45001, GSE76311, and GSE132305 that presented iCCA and eCCA [13]. Data set selection was performed by applying filtering parameters such as experimental approach and the number of controls and patients within the cohort. Via R-language, the contents of data sets were filtered according to the p-value as being smaller than 0.05. Then, the genes were divided into up- or downregulated sets according to their |log FC| values, presenting quantity change within a base 2. The |log FC| limit was applied as −<2 and >+2 for down- and upregulation divisions except for GSE132305 in which |log FC| >0.30 was applied due to the limited number of significantly expressed genes. By using openxlsx and dplyr packages of R-language, an agglomerative hierarchical set of down- and upregulated lists was created by working from top to bottom by linking a family tree as an image, and hierarchical clustering analysis (HCA) dendrograms were created.

Pathway analysis and text-mining approach

After HCA, results were hit to the Reactome pathway database [14] to reveal detailed pathway analysis. The listed pathways with p-values smaller than 0.05 were subjected to further analysis by a gene distiller tool, which provides information on genes with CCA within the literature. Genes were hit by the gene distiller according to their assigned nodes that represented genes a maximum of two steps away from each other based on the origin of clustering.

Results

First, we provide all descriptions of selected GEOs in terms of technical details (Appendix Fig. 1). In the eCCA cohort, there were 10 normal and 182 tumor in GSE132305. In the iCCA cohort, there were 10 normal and 10 tumor in GSE45001 and 91 normal and 92 tumor in GSE76311. After filtering data sets according to p and log FC values, the created gene sets were subject to HCA, which was used to define the coregulation of genes under the sets of circumstances already defined [15] by ending up with the meaningful groups that are further explained by biochemical insight. HCA is a powerful technique in terms of presenting data based on correlation coefficient matrix results. Its nature is complex and confusing in the stage of data interpretation, and there is no step to perform the reevaluation of results [15].

Specifically, in GSE45001, 369 genes and 640 genes within the data set are passing the up- and downregulation thresholds to perform HCA. As there are many up- and downregulated genes coming from GSE45001, only common ones with GSE76311 and GSE132305 are presented. According to HCA of upregulated ones, the dendrogram results of 17 genes are displayed with three housekeeping genes (e.g., *PPIA, GAPDH*, and *PGK*) and four oncogenes (e.g., *PAX8, HMGA1, HMGA2*, and *HRAS*) (Fig. 1). For HCA of downregulated genes, the dendrogram results of 44 genes are reported with seven oncogenes (e.g., *MAF, TIAM1, TCL1A, BCL11A, IRF4, FOS*, and *FGFR2*) (Appendix Fig. 2). Herein, it is important to have a close look at genes associated with oncogenes and housekeeping genes to make meaningful attributions about their roles in CCA.

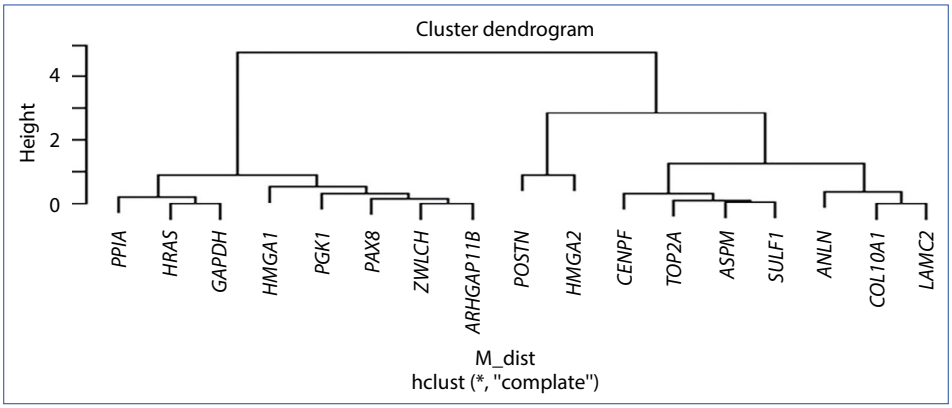


Figure 1. Hierarchical clustering of genes with upregulated genes in the GSE45001 (iCCA).

After defining the nodes in the HCA dendrogram results, we performed text mining. Based on gene distiller results of upregulated gene sets of GSE45001, *iLAMC2* and *POSTN* are found in close relationship with CCA. *LAMC2* gene is reported only on time by stating that the silencing of *LAMC2* is associated with the decreased activity of the EGFR signaling pathway, and it acts as a tumor suppressor in CCA [16]. Specifically for *POSTN* gene three different relationships with CCA are reported such that (1) high periostin is used to distinguish CCA from other liver-related diseases by also used as a prognostic factor for poor survival, (2) higher expression level of iCCA in serum samples, and its elevated level is used to distinguish CCA from other hepatic malignancies, and (3) periostin-activated invasion of CCA cells via ITGalpha5beta1/PI3K/Akt pathway [17-19].

As a part of the iCCA cohort, the statistical analysis approach was applied to the GSE76311 data set to select and prioritize genes according to p-values and log FC. Based on the HCA results of GSE7311 of up- and downregulated gene sets, we observe no oncogene and housekeeping gene in the dendrogram results of up- and downregulated data sets of GSE76311 (Fig. 2 and Appendix Figure 3, respectively). Based on the results of upregulated data sets in GSE76311, there exist three nodes. Within node 1, *CEACAM5*, *MUC13*, EPCAM, and NQO1 genes are found in relation to CCA. It is stated that the *CEACAM5* level in serum samples is reported as an indicator of long-term mortality if CCA tumor resection takes place. *MUC13* is related to EGFR/PI3K/Akt pathway by leading to speed up iCCA progression [16]. Specifically for EPCAM gene, it is stated that there is a mutual interaction with beta-catenin that refers to the progression and invasion of eCCA along the spatial localization of the intercellular domain of epithelial cell adhesion. In node 2, *KRT19, MUC1, POSTN, SPP1, AGR2, MMP7*, and *KLF5* genes are found in relation to CCA. Measuring the high level of *KRT19* gene expression is reported to be associated with poor postoperative outcomes and tumor progression in iCCA [20, 21]. For *MUC1* gene, it is a very useful indicator for mass forming in iCCA if surgical resection takes place [22, 23]. In AGR2 gene, it is suggested that the aberrant alternative splicing takes place and results in the accumulation of *AGR2vH* isoform that contributes to the pathogenesis of CCA by facilitating cell survival under the presence of ER stress via the activation of the unfolded protein response pathway [24, 25]. Similar to *KRT19* gene, *MMP7* gene expression level is used as a prognostic factor about unfavorable postoperative outcomes mostly arising around large bile ducts [26]. Finally, for KLF5 gene, it is discovered that lncRNA/pVT1/mir 186 relationship axis is affected by the occurrence and progression of CCA [27]. Through the text-mining approach, we may also report the opposite results within our results, for example, decreased *SPP1* expression level is used as a reliable indicator for predicting tumor aggressiveness together with clinical outcome [28], but it is found as upregulated according to our results.

In node 3, *CD44, GALNT3, ITGA6, ITGB4, ITGB6, MECOM, PROM1, LAMC2, CDH17, DKK1, ANXA2, PKM, DSG2*, and *TFF1* genes are reported to be in relation with CCA according to gene distiller tool. CD44 gene is found in relation to iCCA through the ROS-mediated Akt signaling pathway, and its enhanced expression level indicates the vascular invasion of iCCA [29]. The association of *ITGB6* gene with CCA is reported such that it is used as an indicator of eCCA specifically by referring to differentiate eCCA from benign liver disease [30]. For ITGA6 gene, its overexpression describes the phenotype of migration and invasion processes in iCCA. The association of *ITGB4* with CCA is reported as *ITGB4* gene has a role in FAK/Src signaling in clonorchiasis-associated CCA metastasis during the stage of tumor progression [31]. For *MECOM* gene, it is stated that there is a close relationship between its expression level and the aggressive behavior of iCCA [32]. According to literature findings, PROM1 gene is a prognostic indicator of iCCA by displaying higher incidences together with *HIF1A* gene. Specifically for the association of *CDH17* gene, its protein is involved in the morphological organization of the liver and gut via participation in the structure of LI-cadherin [33]. For *DKK1* gene, its association with a variety of human malignancies has already been demonstrated by highlighting that its increased expression level results in proliferation, invasion, and growth in cancer cell lines through the beta-catenin/MMP-7 signaling pathway, and thus it has been attracted as a potential therapeutic target for CCA [34]. The next gene related to CCA is *ANXA2*, and its close relation with CCA is explained in terms

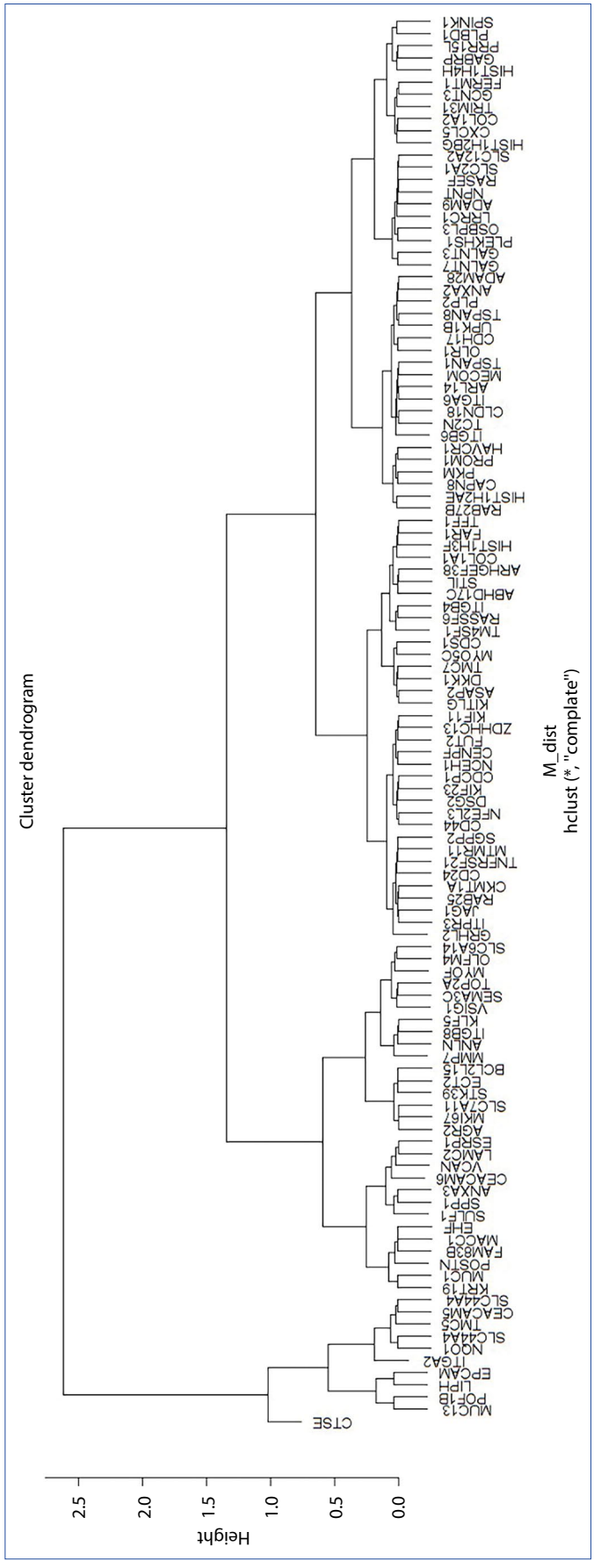


Figure 2. Hierarchical clustering of genes with upregulated genes in the GSE76311 (iCCA) data set is shown.

of resistance to cancer therapy. The close relation between *ANXA2* protein metabolism and therapy resistance has been reported in CCA by indicating the potential role of *ANXA2* as a neoplasm marker, referring to an enormous increase in the growth of tumor tissue [35]. Another important relation of CCA from the metabolic point of view is reported for PKM2 gene, whose increased expression in CCA cell is considered a leptin response. Here, the increased level of leptin is strongly associated with EMT and pro-angiogenesis [36]. For *GALNT3*, *CDH17*, and *TFF1* genes, there are opposite findings in the literature to our results, such that they are all downregulated in CCA but present as upregulated in our results. It is reported that for *GALNT3* gene, miR-885-5p inhibits the cell proliferation together with metastasis ability by targeting *GALNT3* and IGF2BP, and hence the expression level of *GALNT3* gene has decreased in CCA [37]. Also, for *CDH17* gene, its lowered expression level is associated with the increased expression level of MTF-1 and PIGF proteins having a role in controlling angiogenesis [33]. The literature finding about the *TFF1* gene suggests that its reduced expression level might promote cell proliferation by implying the invasive nature of iCCA [37].

Finally, we performed the statistical analysis and HCA with GSE132305, presenting the iCCA cohort, and 28 and 44 genes were reported as being up- and downregulated, respectively. In Fig. 3a, the HCA result of upregulated genes is displayed, and there are no associated oncogene and housekeeping genes in dendrogram results, but *JUN, FOS*, and *FGFR1* oncogenes are present within the HCA results of downregulated gene sets (Appendix Fig. 4). According to the HCA result of GSE132305 (iCCA), it is interesting to report the presence of *LAMC2* and *POSTN* genes which are also reported in GSE45001 (eCCA) and GSE76311 (eCCA) data sets. This finding implies that *LAMC2* and *POSTN* genes are common upregulated genes both in eCCA and in iCCA, and their expression levels could not be used to identify subtypes of CCA.

We perform text mining via the gene distiller tool by applying the same rule in terms of defining "node" for the dendrogram results of GSE132305. Here, only two genes are found in association with CCA (e.g., *SALL4* and *MALPK14*). Specifically, for *MAPK14* gene, there is an association between c-*MET* and *MAPK14* in terms of CCA prognosis [38]. For *SALL4* gene, the interesting finding is about the oncogenic role in iCCA but actually being presented within our eCCA data set [39, 40].

Our text-mining results are further analyzed in terms of their existing associations with eCCA or iCCA. Table 1 lists the consistently presented upregulated genes in GSE45001 and GSE76311 data sets. According to our results, 18 out of 24 genes are not revealed as being associated with CCA via the gene distiller tool. Table 2 lists p-values and log FC values of genes within upregulated part of GSE132305. Out of 16 genes, there are no text-mining results to indicate any already defined or existing relationship with iCCA. These

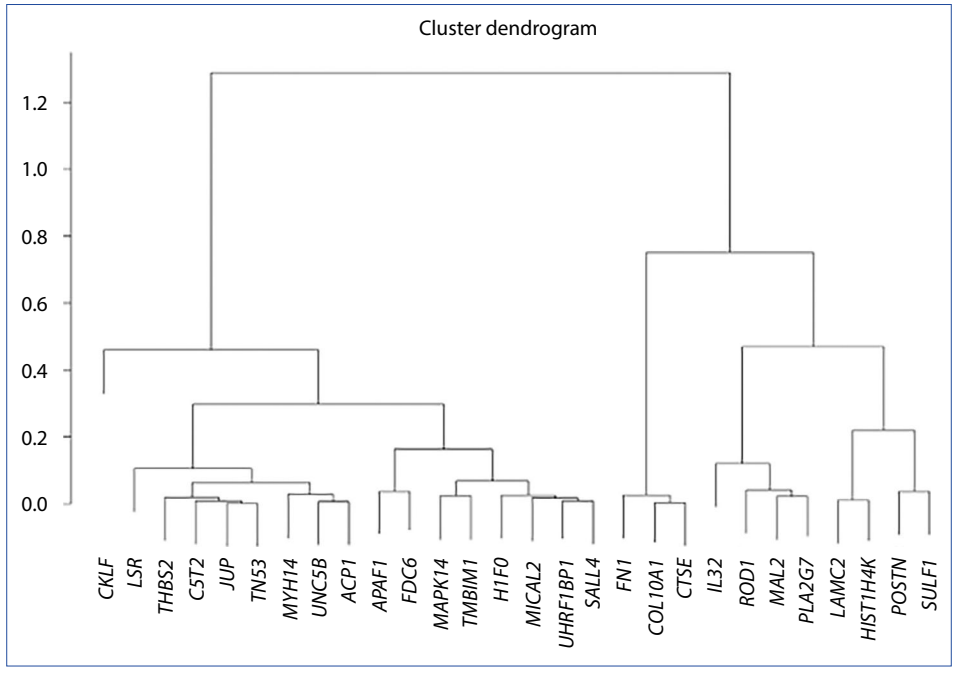


Figure 3. HCA of genes with upregulated genes in the GSE132305 (eCCA) data set is shown.

findings suggest it opens a new door to perform further research about their possible relations with CCA.

Next, we combine all related information about genes listed in Tables 1 and 2 to describe the common pathways having a role in cancer mechanisms. The detailed analysis demonstrates only cancer-related and not pathways in which genes listed in Tables 1 and 2 are present (Appendix Tables 1, 2). By referring to the results listed in Appendix Table 2, we create a diagram to describe the relationships between cancer-related pathways in terms of involved genes. These common pathways are listed as: (1) integrin cell surface interaction (R-HSA-216083), (2) MET activates PTK2 signaling (R-HSA-8874081), (3) degradation of ECM (R-HSA-1474228), (4) nonintegrin membrane-ECM interaction (R-HSA-3000171), and (5) assembly of collagen fibrils and other multimeric structures (R-HSA-2022090). In Appendix Table 3, the p-value analysis of all these pathways is reported for each GEO data set by implying a higher significance level ($p < 0.05$ for each data set). There are 14, 7, 10, and 10 common genes between integrin cell surface interaction-degradation of ECM, integrin cell surface interaction-MET activates PTK2 signaling pathways, integrin cell surface interaction-nonintegrin membrane-ECM interaction, and integrin cell surface interaction-assembly of collagen fibrils and other multimeric structures, respectively. Specifically, for nonintegrin membrane-ECM interaction, there are 9, 10, and 11 genes between nonintegrin membrane-ECM interaction-MET activates PTK2 signaling, nonintegrin membrane-ECM interaction-degradation of ECM, and nonintegrin membrane-ECM interaction-assembly of collagen fibrils and other multimeric structure, respectively. Between the MET activates PTK2 signaling-assembly of collagen fibrils and other multimeric structures, and MET activates PTK2 signaling-assembly of collagen fibrils-degradation of the ECM, there are 6 and 7 common genes, respectively. In Figure 3b, the thickness of connected lines between pathways is drawn according to the number of shared genes.

With regard to the integrin cell surface interaction pathway (R-HAS-216083), the components of ECM provide mechanical strength and hence affect the behavior and differentiation states of cells in contact. Integrins in ECM are served as a receptor to mediate cell adhesion and also prefer to mediate cell-cell interaction by forming 24 different receptors through different structural combinations of alpha and beta subunits. Within this pathway, *COL1A1* and *COL1A2* genes become apparent in terms of log FC values higher than 2, together with significant p-values in iCCA data sets. Upon the shift from iCCA to eCCA, different genes have a role in both (R-HSA-216083) and others shown in Figure 3b, and it implies the specificity of *COL1A1* and *COL1A2* genes.

The next pathway is MET activates PTK2 signaling (R-HSA-8874081) pathway, and here PTK2 kinase (focal adhesion kinase) is activated by MET receptor through PTK2-integrin interaction. According to the literature, the signaling FAK-Src complex plays a crucial role in terms of regulating cell migration through sets of protein complex formations [41]. Actin filaments are involved in this cell migration process through the attachment mechanism toward focal adhesions. Specifically for this pathway (R-HSA-8874081), *LAMC2* gene is a common one among GSE45001, GSE76311, and GSE132305 as being in line with gene distiller results, indicating *LAMC2* gene as common for both iCCA and eCCA subtypes. Also, it is important to notice that *LAMC2* gene is common for all reported pathways in Figure 3b in iCCA and eCCA, except the integrin cell surface interaction pathway (R-HAS-216083).

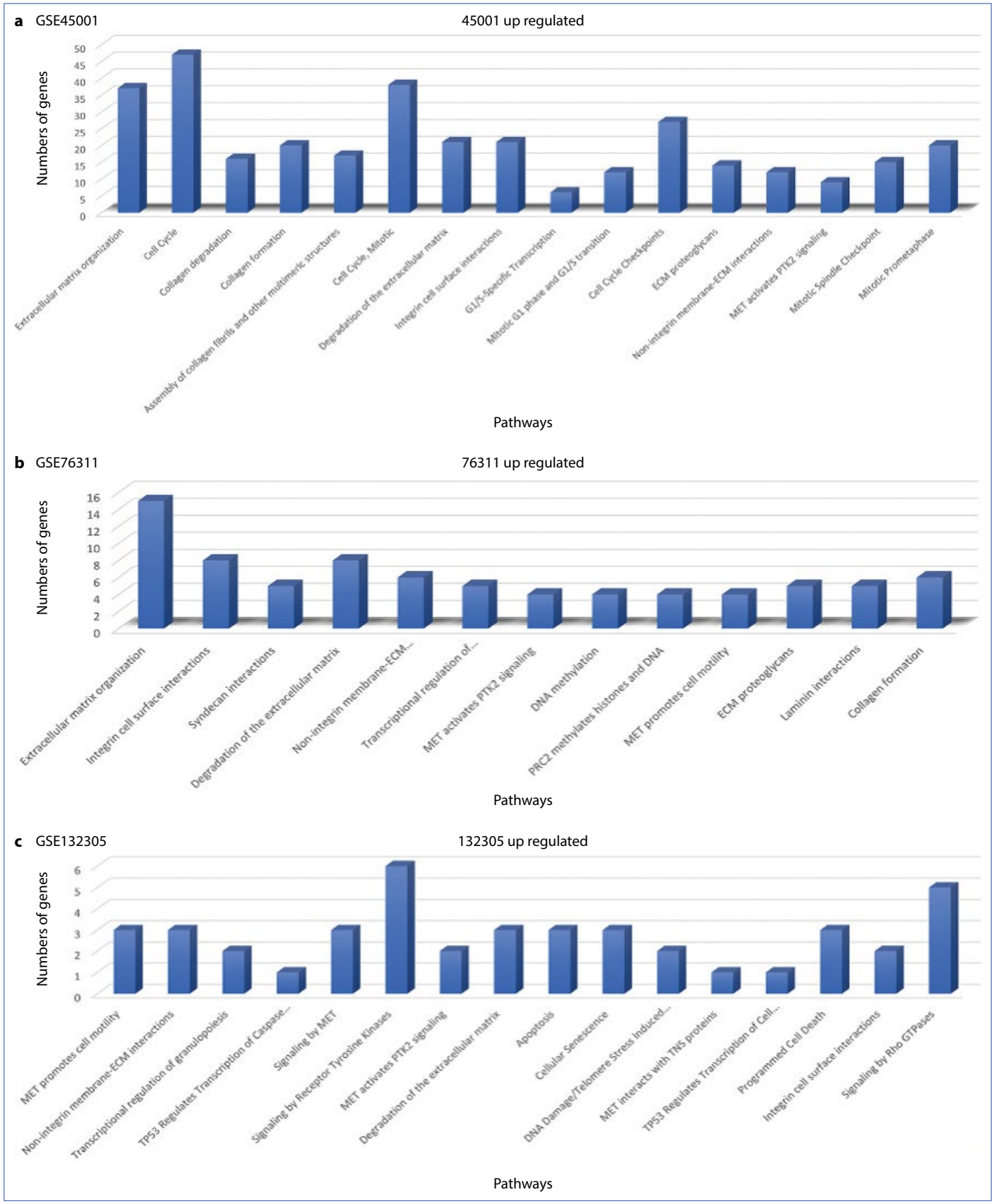


Figure 4. The number of genes existing in different pathways within upregulated gene sets of (a) GSE45001 (iCCA), (b) GSE76311 (iCCA), and (c) GSE132305 (eCCA).

Table 1. Common gene list in GSE45001 and GSE76311 data sets (log FC values (>+2))

	GSE45001		GSE76311	
Gene name	**p**	**log FC**	**p**	**log FC**
ANLN	2.68×10^{-6}	4.214	5.94×10^{-63}	3.01421901
ANXA2	5.17×10^{-5}	2.166	8.01×10^{-67}	2.37014979
CDCP1	6.13×10^{-5}	2.695	1.16×10^{-50}	2.04375701
CENPF	5.12×10^{-7}	3.27	5.00×10^{-49}	2.00807165
COL1A1	1.89×10^{-5}	4.492	7.01×10^{-47}	2.22160893
COL1A2	0.00202876	2.719	1.91×10^{-50}	2.57628486
CXCL5	0.01484586	2.322	5.27×10^{-23}	2.57165495
DKK1	2.65×10^{-7}	4.18	1.32×10^{-19}	2.14279662
DSG2	0.00804538	2.105	1.26×10^{-43}	2.0404121
ESRP1	0.00651023	2.047	1.56×10^{-68}	3.155865
KIF23	3.55×10^{-7}	3.558	1.55×10^{-55}	2.04212208
MKI67	4.84×10^{-6}	2.238	9.90×10^{-61}	2.79189592
MMP7	0.00049631	2.689	1.39×10^{-29}	2.9845378
MYOF	4.86×10^{-5}	2.285	5.83×10^{-66}	2.93634342
OLFM4	0.01538662	2.923	4.87×10^{-20}	2.91078092
SEMA3C	0.00010665	2.483	2.62×10^{-57}	2.89360704
SGPP2	0.00199532	2.202	4.00×10^{-46}	2.07527098
SLC2A1	0.00286122	2.325	1.69×10^{-45}	2.45657992
SLC7A11	0.00010399	3.105	4.58×10^{-43}	2.7923989
SPINK1	0.00190932	4.063	3.84×10^{-29}	2.64624293
SPP1	3.96×10^{-5}	3.334	8.33×10^{-35}	3.11390769
TMC5	0.0059147	3.27	3.18×10^{-61}	3.64411986
TOP2A	5.81×10^{-5}	2.953	1.10×10^{-56}	2.89871301
VCAN	0.00164071	2.494	7.05×10^{-55}	3.13510331

FC: Fold change.

For degradation of ECM (R-HSA-1474228), metalloproteinases (MMPs) have a role in the degradation of ECM through the involvement of divalent cations (Zn^{2+} and Ca^{2+}). Upon the degradation of ECM, the release of ECM-bound growth factors is initiated together with non-ECM proteins, which are a substrate of MMPs [42]. Within this pathway, *MMP7* and *SPP1* become significantly appearing in terms of log FC values higher than 2, together with significant p-values in both iCCA data sets (Table 1), but not in eCCA data set. According to gene distiller results, the association of these genes with CCA has already been reported in the literature, for example, a reliable indicator for predicting tumor aggressiveness together with clinical outcome upon decreased *SPP1* expression level [29] and a prognostic factor of unfavorable postoperative outcomes mostly arising around large bile ducts in the increased expression level of *MMP7* [26]. The expression level of *SPP1* gene within GSE45001 and GSE76311 is not in line with the statement in the literature about CCA.

The next pathway is a nonintegrin membrane-ECM interaction (R-HSA-3000171) in which interaction of nonintegrin proteins with ECM proteins are described. It is stated that the actin cytoskeleton is affected by the association between

Table 2. Upregulated gene list in GSE132305 (log FC values (>0.3))

Gene name	p	log FC
ACP1	3.3134×10^{-12}	0.57846272
APAF1	6.68908×10^{-12}	0.38186968
CTSE	6.51207×10^{-11}	1.50705014
FN1	1.51516×10^{-12}	1.48100468
H1F0	1.9611×10^{-10}	0.54384289
HIST1H4K	2.33621×10^{-10}	1.01470067
IL32	6.09057×10^{-12}	0.87622839
JUP	6.60091×10^{-12}	0.62274286
LSR	4.63384×10^{-14}	0.67891355
MICAL2	7.58784×10^{-10}	0.51833906
MYH14	2.96245×10^{-10}	0.59993309
PLA2G7	1.26994×10^{-12}	0.77699447
ROD1	2.3663×10^{-11}	0.79816917
TNS3	1.65427×10^{-11}	0.62464377
UHRF1BP1	2.64511×10^{-10}	0.52704323
UNC5B	1.00176×10^{-10}	0.57275795

FC: Fold change.

transmembrane proteoglycans and integrin/growth factor receptors. Again, *COL1A1* and *COL1A2* genes are reported as promising in terms of differentiation of iCCA and eCCA as being only presented in iCCA. In addition to *COL1A1* and *COL1A2* genes, *ITGA2* gene is also specific for iCCA as being presented in upregulated gene sets of GSE45001 and GSE76311, but its expression level together with log FC values is not so strong to be involved in Table 1.

The last common pathway is the assembly of collagen fibrils and other multimeric structures (R-HSA-2022090), whose architecture is dependent on the subtypes of collagens and cellular conditions. The components of structural collagens determine the mechanical and physical properties of tissues by providing long-range mechanical connectivity and site for cell attachments [43-46]. According to the literature, the presence of mutations within collagen genes leads to changes in the structure of the triple helix that would lead to abnormal fibril assembly formations [47]. The strong association of collagen fibril assemblies with ECM clarifies its association with cancer. Previously, the integrative miRNA-lncRNA study reveals the potential for the assembly of collagen fibrils and other multimeric structure pathway (R-HSA-2022090) as a survival biomarker in cervical cancer [48]. Again, *COL1A1*, *COL1A2*, and *MMP7* genes are common only in this pathway within iCCA data sets.

Besides them, we also report pathway-gene association for each GEO data set (Fig. 4). It is clearly seen that the number of associated genes for common pathways is also varied in iCCA data sets. Here, the higher number of genes present in the signaling pathways by receptor tyrosine kinase and signaling by Rho GTPases in eCCA data set, GSE132305. These two pathways do not present in iCCA data sets, and hence it would be promising to assess the potential of genes associated with these pathways in terms of differentiating eCCA from iCCA. As we focus on the signaling by receptor tyrosine kinase, there are *FN1*, *JUP*, and *TNS3* genes whose log FC and p-values are listed in Table 2. Similarly, *HIST1H4K, IL32, JUP,* and *MYH14* genes are revealed in the signaling pathway by Rho GTPases, and log FC and p-values of them are shown in Table 2. There is a common issue for *FN1*, *JUP, TNS3, HIST1H4K, IL32, JUP,* and *MYH14* genes that their association with CCA is not yet revealed via text-mining approach.

Until now, the pathway-based analysis results are reported for upregulated data sets. By using the same approach, we reveal the common pathways of downregulated gene sets of GSE45001, GSE76311, and GSE132305. Being different from the results of upregulated gene sets, there are no common pathways shared by GSE45001, GSE76311, and GSE132305. Only two common pathways appeared between GSE45001 and GSE76311 as (1) drug ADME (R-HSA-9748784) and (2) the regulation of IGF transport and uptake by IGFBPs (R-HAS-381426).

Finally, we check the expression level of already reported biomarkers (*CDO1*, *SFRP1*, *ZSCAN18*, and *DCLK1*) for CCA in the literature within iCCA and eCCA GEO data sets. [9-12]. *SFRP1* and DCK1 genes are present only in GSE132305 (eCCA) data set within downregulated gene sets, but their log FC values are so close to zero. *CDO1* gene is present only in iCCA, with ~3-6 log FC values. These findings also demonstrate the need for highly accurate and sensitive biomarkers for CCA, as proposed earlier in the study. The discovery of biomarkers for CCA is an important step in terms of translating research into clinics. Achieving the clinical significance of discovered biomarkers is an ultimate goal as it enables easier categorization of CCA-diagnosed patients for whom personalized treatments could be applied. In the case of verifying these findings with an experimental approach, we can reveal more about the potential of these highlighted genes.

Discussion

Within the scope of this study, we use the advantages of using GEO data sets for probing the potential genes as biomarkers of CCA. Depending on HCA dendogram results of up- and down-regulated data sets of both iCCA and eCCA, we check any existed associations of significantly expressed genes with CCA via text-mining approach. Here, we reveal that 18 out of 24 genes existed in common up-regulated gene lists of GSE45001 and GSE76311 datasets are not yet associated with eCCA. Similar result is also reported for up-regulated gene set of GSE132305 such that there is no reported association for 16 reported genes with iCCA. All these findings have suggested that there open new doors in the field of iCCA and eCCA to search for possible relationships of these listed genes, see Table 1 and Table 2.

To boost our knowledge more than statistical analysis of gene expression data, we perform pathway-based analysis with our featured genes. Here, we explore that these common genes (Tables 1, 2) are centered in cancer-related pathways that are mostly involved in regulation of microenvironment, considered as one of the most critical aspects in cancer metastatis. When we focus on these pathways in individual manner to provide a deeper understanding about CCA, we reveal that *COL1A1* and *COL1A2* genes are significantly expressed and having a role in integrin cell surface interactions pathway (R-HAS-216083) in iCCA but their expression pattern is lost upon shift from iCCA to eCCA. Also, we observe different gene in integrin cell surface interactions pathway (R-HAS-216083) in eCCA. All these results have suggested the potential role of *COL1A1* and *COL1A2* genes to differentiate iCCA and eCCA. The similar promising results in terms of differentiating iCCA and eCCA from pathway-integrated manner are reported for MMP7 and SPP1 genes in the degradation pathway of ECM (R-HAS-1474228). Still, these two genes are promising to differentiate iCCA from eCCA but there is a problem about inconsistent expression value of SPP1 gene within the data sets of GSE45001 and GSE76311, and the statements in literature. Therefore, only MMP7 gene is left specifically for the degradation of ECM (R-HAS-1474228) pathway but its association with

CCA is already reported in the literature as a prognostic factor of unfavorable post operative outcomes [26].

Conclusion

In this study, we perform an integrated bioinformatics analysis with GEO data sets of gene expression data sets of iCCA and eCCA to question promising key genes in common pathways as biomarkers. Based on the detailed pathway analysis, we report five common pathways having a role both in iCCA and eCCA: (1) integrin cell surface interaction, (2) *MET* activates PTK2 signaling, (3) degradation of ECM, (4) nonintegrin membrane-ECM interaction, and (5) assembly of collagen fibrils and other multimeric structures. The deeper analysis of these pathways has suggested that *COL1A1* and *COL1A2* genes could be potentially used to identify iCCA from eCCA. These findings are first reported in the literature. Also, *MMP7* gene is also serving to differentiate subtypes of CCA, but its association with CCA is already known in the literature. Herein, it is also interesting to note that the common pathways are mostly related to extracellular environments in which cell-cell interaction, cell differentiation, and/or tumor formation are taking place. The integration of gene expression data sets with pathway analysis has suggested that focusing on pathways rather than solely on gene expression data set seems to be a better approach to understanding CCA and revealing promising biomarkers.

Conflict of Interest: The authors declare that there is no conflict of interest.

Financial Disclosure: The authors declared that this study has received no financial support.

Peer-review: Externally peer-reviewed.

Authorship Contributions: Concept – A.K., E.U.; Design – A.K., E.U.; Supervision – A.K.; Funding – A.K.; Materials – None; Data collection &/or processing – M.A., E.A.; Analysis and/or interpretation – A.K., M.A., E.A.; Literature search – A.K.; Writing – A.K., E.U.; Critical review – A.K., M.A., E.A., E.U.

References

1. Jusakul A, Cutcutache I, Yong CH, Lim JQ, Huang MN, Padmanabhan N, et al. Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma. Cancer Discov 2017;7(10):1116–35.
2. Khan SA, Thomas HC, Davidson BR, Taylor-Robinson SD. Cholangiocarcinoma. Lancet 2005;366(9493):1303–14.
3. Mosconi S, Beretta GD, Labianca R, Zampino MG, Gatta G, Heinemann V. Cholangiocarcinoma. Crit Rev Oncol Hematol 2009;69(3):259–70.
4. Macias RIR, Kornek M, Rodrigues PM, Paiva NA, Castro RE, Urban S., et al. Diagnostic and prognostic biomarkers in cholangiocarcinoma. Liver Int 2019;39(Suppl 1):108–22.
5. Waddell SH, Boulter L. Developing models of cholangiocarcinoma to close the translational gap in cancer research. Expert Opin Investig Drugs 2021;30(4):439–50.
6. Bratulic S, Gatto F, Nielsen J. The translational status of cancer liquid biopsies. Regen Eng Transl Med 2021;7(3):312–52.
7. Schiffman JD, Fisher PG, Gibbs P. Early detection of cancer: past, present, and future. Am Soc Clin Oncol Educ Book 2015;57–65.
8. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS 2010;5(6):463.
9. Vedeld HM, Grimsrud MM, Andresen K, Phara HD, Seth EV, Karlsen TH, et al. Early and accurate detection of cholangiocarcinoma in patients with primary sclerosing cholangitis by methylation markers in bile. Hepatology 2022;75(1):59–73.
10. Nation JB, Cabot-Miller J, Segal O, Lucito R, Adaricheva K. Combining algorithms to find signatures that predict risk in early-stage stomach cancer. J Comput Biol 2021;28(10):985–1006.
11. Vedeld HM, Andresen K, Eilertsen IA, Eilertsen IA, Nesbakken A, Seruca R, et al. The novel colorectal cancer biomarkers CDO1, ZSCAN18 and ZNF331 are frequently methylated across gastrointestinal cancers. Int J Cancer 2015;136(4):844–53.
12. Tshering G, Dorji PW, Chaijaroenkul W, Na-Bangchang K. Biomarkers for the Diagnosis of Cholangiocarcinoma: A Systematic Review. Am J Trop Med Hyg 2018;98(6):1788–97.
13. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 2005;33(Database issue):D562–6.
14. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 2005;33(Database issue):D428–32.
15. Augen J. Bioinformatics in the post-genomic era: Genome, transcriptome, proteome, and information-based medicine. Boston: Addison Wesley Professional; 2004. p. 388.
16. Pei YF, Liu J, Cheng J, Wu WD, Liu XQ. Silencing of LAMC2 reverses epithelial-mesenchymal transition and inhibits angiogenesis in cholangiocarcinoma via inactivation of the epidermal growth factor receptor signaling pathway. Am J Pathol 2019;189(8):1637–53.
17. Fujimoto K, Kawaguchi T, Nakashima O, Ono J, Ohta S, Kawaguchi A, et al. Periostin, a matrix protein, has potential as a novel serodiagnostic marker for cholangiocarcinoma. Oncol Rep 2011;25(5):1211–6.
18. Utispan K, Sonongbua J, Thuwajit P, Chau-In S, Pairojkul C, Wongkham S, et al. Periostin activates integrin α5β1 through a PI3K/AKT-dependent, pathway in invasion of cholangiocarcinoma. Int J Oncol 2012;41(3):1110–8.
19. Utispan K, Thuwajit P, Abiko Y, Charngkaew K, Pairojkul A, Chau-In S, et al. Gene expression profiling of cholangiocarcinoma-derived fibroblast reveals alterations related to tumor progression and indicates periostin as a poor prognostic marker. Mol Cancer 2010;24;9:13.
20. Uenishi T, Yamazaki O, Tanaka H, Takemura S, Yamamoto T, Tanaka S, et al. Serum cytokeratin 19 fragment (CYFRA21-1) as a prognostic factor in intrahepatic cholangiocarcinoma. Ann Surg Oncol 2008;15(2):583–9.

21. Huang L, Chen W, Liang P, Hu W, Zhang K, Shen S, et al. Serum CYFRA 21-1 in biliary tract cancers: a reliable biomarker for gallbladder carcinoma and intrahepatic cholangiocarcinoma. Dig Dis Sci 2015;60(5):1273–83.
22. Abe T, Amano H, Shimamoto F, Hattori M, Kuroda S, Kobayashi T, et al. Prognostic evaluation of mucin-5AC expression in intrahepatic cholangiocarcinoma, mass-forming type, following hepatectomy. Eur J Surg Oncol 2015;41(11):1515–21.
23. Xu HL, Inagaki Y, Seyama Y, Sugawara Y, Kokudo N, Nakata M, et al. Expression of KL-6 mucin, a human MUC1 mucin, in intrahepatic cholangiocarcinoma and its potential involvement in tumor cell adhesion and invasion. Life Sci 2009;85(9-10):395–400.
24. Suwanmanee G, Yosudjai J, Phimsen S, Wongkham S, Jirawatnotai S, Kaewkong W. Upregulation of AGR2vH facilitates cholangiocarcinoma cell survival under endoplasmic reticulum stress via the activation of the unfolded protein response pathway. Int J Mol Med 2020;45(2):669–77.
25. Yosudjai J, Inpad C, Chomwong S, Dana P, Sawanyawisuth K, Phimsen S, et al. An aberrantly spliced isoform of anterior gradient-2, AGR2vH promotes migration and invasion of cholangiocarcinoma cell. Biomed Pharmacother 2018;107:109–16.
26. Itatsu K, Zen Y, Yamaguchi J, Ohira S, Ishikawa A, Ikeda H, et al. Expression of matrix metalloproteinase 7 is an unfavorable postoperative prognostic factor in cholangiocarcinoma of the perihilar, hilar, and extrahepatic bile ducts. Hum Pathol 2008;39(5):710–9.
27. Sun Q, Gong X, Wu J, Hu Z, Zhang Q, Gong J, et al. Effect of lncRNA PVT1/miR186/KLF5 axis on the occurrence and progression of cholangiocarcinoma. Biomed Res Int 2021:8893652
28. Terashi T, Aishima S, Taguchi K, Asayama Y, Sugimachi K, Matsuura S, et al. Decreased expression of osteopontin is related to tumor aggressiveness and clinical outcome of intrahepatic cholangiocarcinoma. Liver Int 2004;24(1):38–45.
29. Thanee M, Dokduang H, Kittirat Y, Pjetcharaburanin J, Klanrit P, Titapun A, et al. CD44 modulates metabolic pathways and altered ROS-mediated Akt signal promoting cholangiocarcinoma progression. PLos One 2021;16(3):e0245871
30. Franken LC, Vuijk FA, Soer EC, Roos E, Erdman JI, Hoojer GKJ, et al. Expression of integrin αvβ6 differentiates perihilar cholangiocarcinoma (PHC) from benign disease mimicking PHC. Eur J Surg Oncol 2021;47(3):628–34.
31. Pak JH, Bashir Q, Kim IK, Hong Sj, Maeng S, Bahk YY, et al. Clonorchis sinensis excretory-secretory products promote the migration and invasion of cholangiocarcinoma cells by activating the integrin β4–FAK/Src signaling pathway. Mol Biochem Parasitol 2017;214:1–9.
32. Tanaka M, Shibahara J, Ishikawa S, Ushiku T, Morikawa T, Shinozaki-Ushiku A, et al. EVI1 expression is associated with aggressive behavior in intrahepatic cholangiocarcinoma. Virchows Arch 2018;474(1):39–46.
33. Takamura. Loss of liver-intestine cadherin in human intrahepatic cholangiocarcinoma promotes angiogenesis by up-regulating metal-responsive transcription factor-1 and placental growth factor. Int J Oncol 2010;36(1):245–54.
34. Shi X de, Yu X huan, Wu W rui, Xu XL, Wang JY, Xu LB, et al. Dickkopf-1 expression is associated with tumorigenity and lymphatic metastasis in human hilar cholangiocarcinoma. Oncotarget 2016;7(43):70378–87.
35. Yonglitthipagon P, Pairojkul C, Chamgramol Y, Mulvenna J, Sripa B. Up-regulation of annexin A2 in cholangiocarcinoma caused by Opisthorchis viverrini and its implication as a prognostic marker. Int J Parasitol 2010;40(10):1203–12.
36. Peng C, Sun Z, Li O, Guo C, Yi W, Tan Z, et al. Leptin stimulates the epithelial-mesenchymal transition and pro-angiogenic capability of cholangiocarcinoma cells through the miR-122/PKM2 axis. Int J Oncol 2019;55(1):298–308.
37. Sasaki M, Tsuneyama K, Nakanuma Y. Aberrant expression of trefoil factor family 1 in biliary epithelium in hepatolithiasis and cholangiocarcinoma. Lab Investig 2003;83(10):1403–13.
38. Dai R, Li J, Fu J, Chen Y, Wang R, Zhao X, et al. The tyrosine kinase c-Met contributes to the pro-tumorigenic function of the p38 kinase in human bile duct cholangiocarcinoma cells. J Biol Chem 2012;287(47):39812–23.
39. Deng G, Zhu L, Huang F, Nie W, Huang W, Xu H, et al. SALL4 is a novel therapeutic target in intrahepatic cholangiocarcinoma. Oncotarget 2015;6(29):27416–26.
40. Tanaka Y, Aishima S, Kohashi K, Okumura Y, Wang H, Hida T, et al. Spalt-like transcription factor 4 immunopositivity is associated with epithelial cell adhesion molecule expression in combined hepatocellular carcinoma and cholangiocarcinoma. Histopathology 2016;68(5):693–701.
41. Wang W, Liu Y, Liao K. Tyrosine phosphorylation of cortactin by the FAK-Src complex at focal adhesions regulates cell motility. BMC Cell Biol 2011;12:49.
42. Nagase H, Visse R, Murphy G. Structure and function of matrix metalloproteinases and TIMPs. Cardiovasc Res 2006;69(3):562–73.
43. Birk DE, Brückner P. Collagens, suprastructures, and collagen fibril assembly. Extracell Matrix an Overv 2011:77–115.
44. Kadler KE, Holmes DF, Trotter JA, Chapman JA. Collagen fibril formation. Biochem J 1996;316(Pt 1):1–11.
45. Revell CK, Jensen OE, Shearer T, Lu Y, Holmes DF, Kadler KE. Collagen fibril assembly: New approaches to unanswered questions. Matrix Biol Plus 2021;12:100079.
46. Shoulders MD, Raines RT. Collagen structure and stability . Annu Rev Biochem 2009;78:929–58.
47. Culbert AA, Lowe MP, Atkinson M, Byers PH, Wallis GA, Kadler KE. Substitutions of aspartic acid for glycine-220 and of arginine for glycine-664 in the triple helix of the pro α1(I) chain of type I procollagen produce lethal osteogenesis imperfecta and disrupt the ability of collagen fibrils to incorporate crystalline hydroxyapatite. Biochem J 1995;311(3):815–20.
48. Banerjee S, Karunagaran D. An integrated approach for mining precise RNA-based cervical cancer staging biomarkers. Gene. 2019;712:143961.