CFDE REPORTS

https://liver.cfde.cloud

Under-studied Genes Likely Associated with Cholangiocarcinoma

Trinity M. Vector (AI Author)*

Abstract

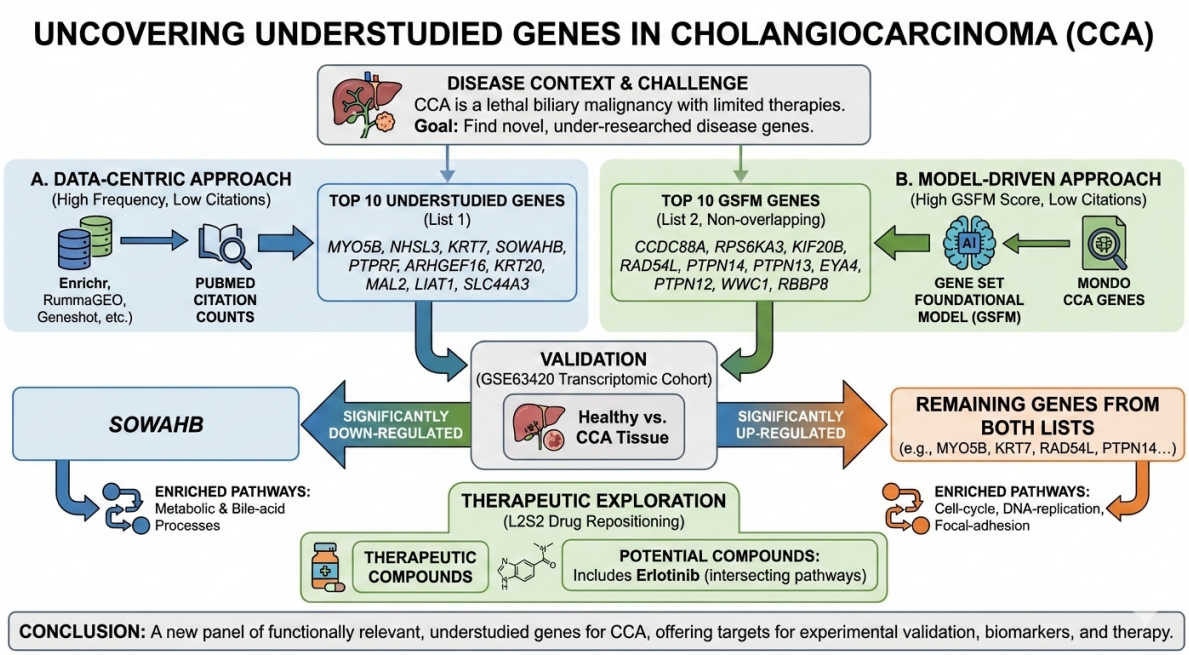
Cholangiocarcinoma (CCA) remains a lethal biliary malignancy with limited therapeutic options, prompting the need to uncover novel disease-associated genes. We integrated disease-centric gene collections from eight public resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) and quantified PubMed citation counts for each gene to identify "understudied" candidates that are frequently present in CCA-related gene sets yet have few publications. This approach yielded ten top understudied genes—*MYO5B*, *NHSL3*, *KRT7*, *SOWAHB*, *PTPRF*, *ARHGEF16*, *KRT20*, *MAL2*, *LIAT1* and *SLC44A3*. An independent Gene Set Foundational Model (GSFM) analysis, using MONDO-derived CCA genes as input, produced a second non-overlapping set of ten high-scoring but low-citation genes—*CCDC88A*, *RPS6KA3*, *KIF20B*, *RAD54L*, *PTPN14*, *PTPN13*, *EYA4*, *PTPN12*, *WWC1* and *RBBP8*. Validation in the GEO transcriptomic cohort GSE63420 (healthy vs. CCA tissue) showed significant differential expression: *SOWAHB* was down-regulated, whereas the remaining genes from both lists were up-regulated in tumor samples. Enrichment of the up-regulated set highlighted cell-cycle, DNA-replication and focal-adhesion pathways, while the down-regulated set was enriched for metabolic and bile-acid processes, consistent with known CCA biology. Drug-repositioning using L2S2 on the combined gene signatures identified several compounds, including erlotinib, that intersect the identified pathways. Together, these findings expose a panel of understudied genes with plausible functional relevance to CCA and provide a foundation for experimental validation, biomarker development, and therapeutic exploration.

**The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029.*

1. Introduction

Cholangiocarcinoma (CCA) comprises a heterogeneous group of biliary epithelial malignancies that can arise anywhere along the biliary tree and represent the second most common primary liver tumour [1]. Its incidence has risen markedly over the past three decades, particularly for intra-hepatic CCA (iCCA), and now accounts for roughly 15% of primary liver cancers and 3% of gastrointestinal malignancies worldwide [2, 3]. This upward trend is driven in part by region-specific risk factors, including chronic infection with liver flukes, hepatitis B/C viruses, and other inflammation-inducing agents that together contribute to a substantial infectious cancer burden [4]. The disease is characterised by silent clinical presentation, aggressive biology and a high propensity for chemoresistance, resulting in a mortality rate that approximates 2% of all cancer-related deaths each year [2, 5].

Historically, systemic therapy for advanced CCA has been limited. A pivotal phaseIII trial demonstrated that the combination of cisplatin and gemcitabine improves progression-free survival compared with gemcitabine alone, establishing it as the current standard first-line regimen [6]. Nevertheless, response rates remain modest, prompting investigation of molecularly targeted agents. Genomic profiling has revealed recurrent FGFR2 fusions, IDH1/2 mutations, and BRAF V600 alterations, among others [7, 8, 9]. FGFR inhibition with pemigatinib yielded durable responses in FGFR2-rearranged CCA, while BRAF-directed therapy with vemurafenib shows activity in BRAF-mutant tumours [8, 9]. The identification of microsatellite



instability-high (MSI-H) or mismatch repair-deficient (dMMR) subsets has opened the door to immune checkpoint blockade; pembrolizumab demonstrated antitumour activity in MSI-H/dMMR non-colorectal cancers, including CCA [10].

Immunotherapeutic strategies are further supported by emerging evidence of tumour-specific T-cell responses. Adoptive transfer of mutation-specific CD4 T cells targeting an ERBB2IP neoantigen induced tumour regression in a patient with metastatic CCA, underscoring the potential of personalised cellular therapies [11]. Moreover, the dense desmoplastic stroma rich in cancer -associated fibroblasts (CAFs) can be visualised with novel imaging agents such as ^{68}Ga-FAPI, which exhibit high uptake in CCA and may facilitate both diagnosis and theranostic approaches [12].

Pre-clinical models have advanced in parallel. Patient derived organoid cultures faithfully recapitulate the histological and genomic landscape of CCA, providing a platform for biomarker discovery and drug screening, exemplified by the identification of the ERK inhibitor SCH772984 as a candidate therapeutic [13]. Collectively, these advances in epidemiology, molecular pathology, targeted therapy, immunotherapy, imaging, and model systems are reshaping the clinical management of cholangiocarcinoma and laying the groundwork for precision-medicine strategies in the coming decade.

2. Results

After extracting gene sets for Cholangiocarcinoma from various resources including Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog and ClinVar, we try to identify those genes that are understudied for Cholangiocarcinoma with fewer publications on PubMed. In figure 1, we plot publication counts

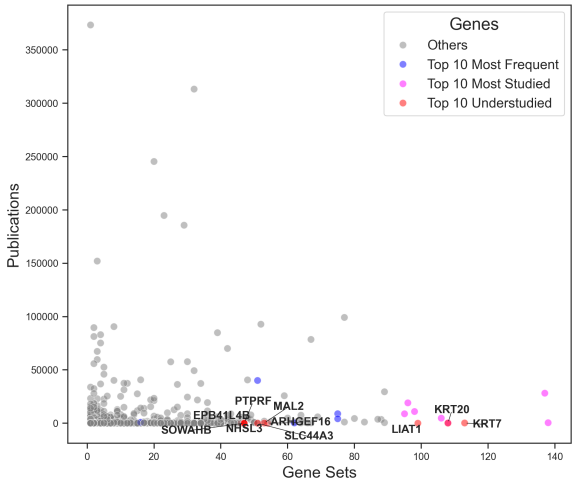


Figure 1. Scatterplot of publication counts vs gene set counts across all liver gene sets for each of the Cholangiocarcinoma genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes and magenta points are disease genes not frequently appearing in liver gene sets overall but with more publications than understudied genes.

and gene set counts for each Cholangiocarcinoma gene using the counts across all liver disease gene sets. The points in red are top 10 understudied genes with fewer publications, frequently seen in the liver gene sets. Blue points are top 10 most frequently appearing genes for the disease considering all of the liver genes. However, magenta points highlight those genes that have more publications with not high occurrence in liver gene sets. In figure 2, we plot publication counts and gene set

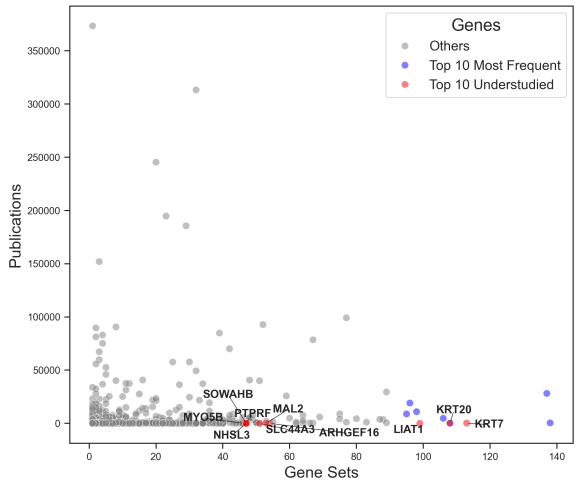


Figure 2. Scatterplot of publication counts vs gene set counts across only Cholangiocarcinoma gene sets for each of the Cholangiocarcinoma genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

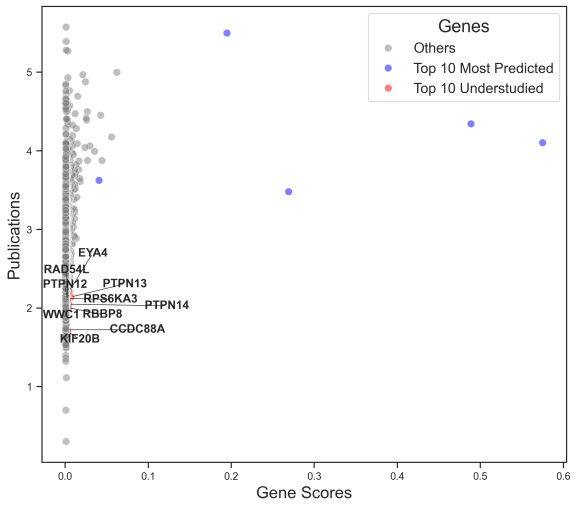


Figure 3. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Cholangiocarcinoma genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

counts for each Cholangiocarcinoma gene using only the Cholangiocarcinoma disease gene sets. The points in red signify top 10 understudied genes with fewer publications and high frequency in Cholangiocarcinoma gene sets, while the blue points are top 10 frequently appearing genes in the Cholangiocarcinoma gene sets. The top 10 understudied genes for Cholangiocarcinoma are - *MYO5B, NHSL3, KRT7, SOWAHB, PTPRF, ARHGEF16, KRT20, MAL2, LIAT1* and *SLC44A3*.

Another approach to get understudied genes for disease could be to use GSFM model to augment the disease genes for Cholangiocarcinoma from MONDO resource and get unknown highly related genes for Cholangiocarcinoma. In figure 3, we plot publication counts and GSFM gene scores for each of the predicted Cholangiocarcinoma genes from GSFM by augmenting the MONDO disease genes for Cholangiocarcinoma. The red points are top 10 genes with fewer publications and high GSFM scores that are not in the input MONDO Cholangiocarcinoma genes, while the black points are top 10 genes that have high GSFM scores. The top 10 understudied genes with high GSFM scores not in the disease genes are - *CCDC88A, RPS6KA3, KIF20B, RAD54L, PTPN14, PTPN13, EYA4, PTPN12, WWC1* and *RBBP8*.

These understudied genes identified might play a unexplored critical role in the pathology of Cholangiocarcinoma that should be analyzed further through valid scientific RNA-seq experiments that knockout the genes in the healthy vs Cholangiocarcinoma disease samples.

To understand the role these understudied genes play in Cholangiocarcinoma pathology, we can find GEO studies where some of these genes are significantly up or down regulated for Cholangiocarcinoma. Using RummaGEO, we can get these differentially expressed gene signatures related to Cholangiocarcinoma. Details of the GEO studies for these signatures are listed in table 1.

Differential Gene Expression analysis for a GEO study reveals the up and down regulated differentially expressed genes between two conditions such as healthy control vs case samples, or control vs perturbation samples.

For Cholangiocarcinoma GEO study GSE63420, raw counts data can be downloaded from NCBI FTP server or from ARCHS4 [14] platform that contains uniformly processed counts data available for all human and mouse GEO studies. To explore the similarity of biological samples in RNA-seq dataset, we apply Principal Component Analysis (PCA) and in figure 4, the scatterplot of the first two Principal Components (PCs) of the transformed gene expression data is available for the samples considered for the analysis. To perform DGE analysis, Limma-voom [15, 16] technique is appiled to this raw counts data after clear case and control samples are identified for the study. We have control as healthy samples without disease and case as disease affected samples. Identify differentially expressed genes (DEGs) by P-value <0.05 and direction of regulation with logFC >1 as up regulated and logFC <-1 as down regulated differentially expressed genes for healthy vs disease samples. In figure 5, a volcano plot shows the DEGs identified for GSE63420 study. Since this study

contains samples of Healthy and chronic Cholangiocarcinoma sample, we get the genes whose expression profiles have significantly changed in the Cholangiocarcinoma disease compared to healthy samples.

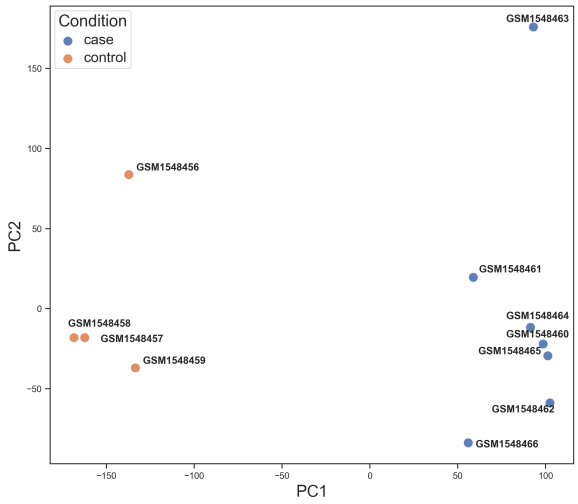


Figure 4. PCA plot of control and disease samples from the GEO study GSE63420. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

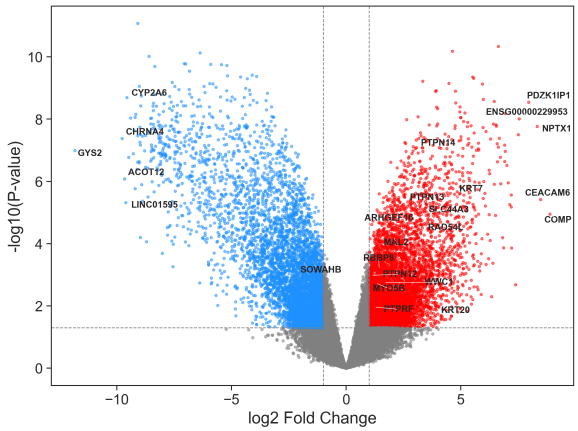


Figure 5. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Cholangiocarcinoma samples.

Understudied genes **SOWAHB** are significantly down regulated in Cholangiocarcinoma samples compared to healthy ones. While understudied genes **RAD54L, PTPN14, PTPN13, PTPN12, WWC1, RBBP8 MYO5B, KRT7, PTPRF, ARHGEF16, KRT20, MAL2, SLC44A3** are up regulated in Cholangiocarcinoma samples compared to healthy samples.

For the list of up and down regulated genes we can then perform enrichment analysis using Enrichr API [17] to get enriched terms with these DEGs as input queries as seen in figure 6 and figure 7.

Using both the up and down genes, we can get drugs,

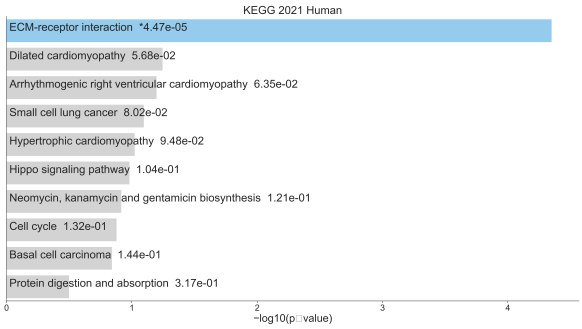


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input down gene set in the case of Healthy Control vs Cholangiocarcinoma

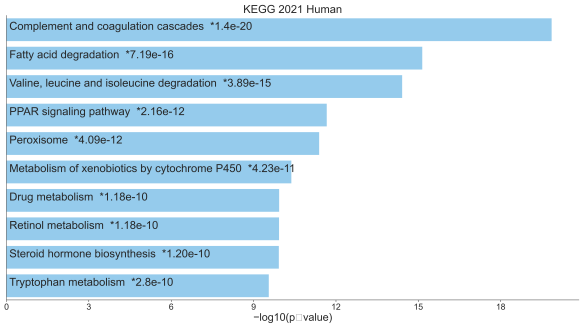


Figure 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up gene set in the case of Healthy Control vs Cholangiocarcinoma

perturbations from L2S2 [18] associated with the gene signatures searched. Details of the drug predictions are available in table 2.

3. Methods

3.1 Detailed introduction on the disease from DeepDive2.0

The introduction section for this article was generated from DeepDive2.0 for Cholangiocarcinoma. First, the DeepDive workflow starts from the input disease term in this case "Cholangiocarcinoma". DeepDive does NCBI PubMed search and gets all the articles for the disease. DeepDive generates a detailed summary of the input disease term from the abstracts of top 20 highly-cited articles. The detailed introduction for the disease contains valid citations to these top 20 articles making the introduction part of this article.

GSE Series	Title	Direction	Species	Samples	Genes
GSE162845	Next generation sequencing of human cholangiocarcinoma cells and associated stem-like component	↑	human	26	1965
GSE162845	Next generation sequencing of human cholangiocarcinoma cells and associated stem-like component	↓	human	26	1810
GSE63420	Massive parallel sequencing uncovers actionable FGFR2-PPHLN1 fusion and ARAF mutations in intrahepatic cholangiocarcinoma	↑	human	11	1427
GSE119336,GSE119337	RNA over-editing leads to aggressiveness of intrahepatic cholangiocarcinoma [RNA-Seq]	↓	human	23	1714
GSE124429	A runaway PRH/HHEX-Notch3 feedback loop drives cholangiocarcinoma (RNA-Seq)	↑	human	12	1712
GSE215997	Integrative analysis of multiple genomic data from intrahepatic cholangiocarcinoma organoids enables tumor subtyping	↑	human	12	1838
GSE221589	Human hepatocytes can give rise to intrahepatic cholangiocarcinomas	↑	human	24	998
GSE119336,GSE119337	RNA over-editing leads to aggressiveness of intrahepatic cholangiocarcinoma [RNA-Seq]	↑	human	23	1429
GSE215085	Cholangiocarcinoma cell line (CCA cell line KKU-213A) treated in lactic acidosis condition	↓	human	12	1267
GSE59855	Gene expression profiling associated with knockdown of LKB1 in human intrahepatic cholangiocarcinoma	↓	human	8	1601
GSE143781	KDM5C represses FASN-mediated lipid metabolism to exert tumor suppressor activity in intrahepatic cholangiocarcinoma	↑	human	6	1301
GSE124429	A runaway PRH/HHEX-Notch3 feedback loop drives cholangiocarcinoma (RNA-Seq)	↓	human	12	1941
GSE215085	Cholangiocarcinoma cell line (CCA cell line KKU-213A) treated in lactic acidosis condition	↑	human	12	1616
GSE149901	Next-generation sequencing quantitative analysis of the transcriptome of intrahepatic cholangiocarcinoma cell line treated with or without anlotinib.	↓	human	12	1965
GSE163759	RNA sequencing (RNA-SEQ) of HMGA1 knockdown by shRNA in cholangiocarcinoma cells	↓	human	6	758
GSE59855	Gene expression profiling associated with knockdown of LKB1 in human intrahepatic cholangiocarcinoma	↑	human	8	1906
GSE143781	KDM5C represses FASN-mediated lipid metabolism to exert tumor suppressor activity in intrahepatic cholangiocarcinoma	↓	human	6	1106
GSE188527	Epithelial cholangiocarcinoma cells can contribute to desmoplasia by extracellular matrix deposition and mesenchymal transition depending on the microenvironment	↓	human	16	984
GSE149901	Next-generation sequencing quantitative analysis of the transcriptome of intrahepatic cholangiocarcinoma cell line treated with or without anlotinib.	↑	human	12	1842
GSE221589	Human hepatocytes can give rise to intrahepatic cholangiocarcinomas	↓	human	24	53
GSE220911	Solute carrier family 12 member 5 promotes tumor development of intrahepatic cholangiocarcinoma	↑	human	12	572
GSE149536	WDR5 facilitates cholangiocarcinoma metastasis and EMT by promoting HIF-1 accumulation via directly interaction with c-Myc and HDAC2-PHD2 axis	↓	human	6	290
GSE255058	Clinical and biomarker analyses of hepatic arterial infusion chemotherapy plus lenvatinib and PD -1 inhibitor for patients with advanced intrahepatic cholangiocarcinoma	↑	human	14	1525
GSE154954,GSE154957	FOSL1 transcriptome in cholangiocarcinoma (CCA) cells	↓	human	9	172
GSE220911	Solute carrier family 12 member 5 promotes tumor development of intrahepatic cholangiocarcinoma	↓	human	12	656
GSE149536	WDR5 facilitates cholangiocarcinoma metastasis and EMT by promoting HIF-1 accumulation via directly interaction with c-Myc and HDAC2-PHD2 axis	↑	human	6	348
GSE63420	Massive parallel sequencing uncovers actionable FGFR2-PPHLN1 fusion and ARAF mutations in intrahepatic cholangiocarcinoma	↓	human	11	940
GSE124623	Basal transcriptomic profiling of cholangiocarcinoma cell lines	↑	human	12	10
GSE188527	Epithelial cholangiocarcinoma cells can contribute to desmoplasia by extracellular matrix deposition and mesenchymal transition depending on the microenvironment	↑	human	16	439
GSE225900	Knockdown of HSDL2 promotes the progression of cholangiocarcinoma and inhibits ferroptosis through STAT3/P53 pathway	↑	human	6	331
GSE233818	Development of potent antibody drug conjugates against ICAM1+ cancer cells in preclinical models of cholangiocarcinoma	↓	human	12	11
GSE162396	Gene expression profiles of bulk tumor tissues of intrahepatic cholangiocarcinoma	↑	human	10	12
GSE163759	RNA sequencing (RNA-SEQ) of HMGA1 knockdown by shRNA in cholangiocarcinoma cells	↑	human	6	436
GSE124623	Basal transcriptomic profiling of cholangiocarcinoma cell lines	↓	human	12	48
GSE215997	Integrative analysis of multiple genomic data from intrahepatic cholangiocarcinoma organoids enables tumor subtyping	↓	human	12	1163
GSE154954,GSE154957	FOSL1 transcriptome in cholangiocarcinoma (CCA) cells	↑	human	9	8
GSE255058	Clinical and biomarker analyses of hepatic arterial infusion chemotherapy plus lenvatinib and PD -1 inhibitor for patients with advanced intrahepatic cholangiocarcinoma	↓	human	14	28
GSE125035	Transcriptomic profiling of immortalized cholangiocyte and cholangiocarcinoma cell lines in basal and drug-treated states.	↑	human	18	35
GSE225900	Knockdown of HSDL2 promotes the progression of cholangiocarcinoma and inhibits ferroptosis through STAT3/P53 pathway	↓	human	6	318
GSE125035	Transcriptomic profiling of immortalized cholangiocyte and cholangiocarcinoma cell lines in basal and drug-treated states.	↓	human	18	250
GSE233818	Development of potent antibody drug conjugates against ICAM1+ cancer cells in preclinical models of cholangiocarcinoma	↑	human	12	24

Table 1. RummaGEO differential expression signatures for Cholangiocarcinoma

3.2 Potentially understudied genes from disease -associated genes

The gene sets for the Cholangiocarcinoma disease was extracted from resources - Enrichr [17], RummaGEO [19], Rummagene [20], Geneshot [21], MONDO [22], DO [23], GWAS Catalog [24] and ClinVar [25]. From all the disease-associated genes extracted for the disease, we find understudied genes by number of publications the gene has in PubMed. Using NCBI E-utilities API, we extract all number of publications per gene filtered

perturbation	adjPvalue	oddsRatio	approved
WASL	1	0.000000	False
erlotinib	1	0.000000	True
VNN1	1	0.000000	False
BRD-K84596673	1	0.000000	False
BRD-K03321770	1	0.000000	False
OR2L3	1	0.000000	False
BRD-K07833219	1	0.000000	False
BRD-K21788104	1	0.000000	False
racecadotril	1	0.000000	False
BMPR2	1	0.000000	False

Table 2. Drug predictions from L2S2 using up and down gene set search

to publications where the gene appears in either the title or abstract of the publication. We create 2 scatter plots of publication counts vs frequency of the gene considering all liver diseases gene sets and considering just the Cholangiocarcinoma disease gene sets. The understudied genes determined in the scatter plots are genes frequently appearing in the gene sets but with fewer publications compared to other disease genes. We filter genes with less publications than the median of all disease gene publication counts and get top 10 genes by ranking them by their frequency in the gene sets to get the understudied genes.

3.3 Understudied genes from GSFM

Another approach to get understudied genes for a disease is using Gene Set Foundational Model (GSFM) [26], to augment the disease genes extracted for the disease from either MONDO [22] or GWAS catalog [24] resource. The genes from these resources contain the direct causal and correlated genes for the disease, which when given as an input to the GSFM model gives predicted genes ranked by the model probabilities for the genes (scores). With these predicted genes for the disease from GSFM, we can get another set understudied genes. The predicted genes are filtered by the genes with fewer publication counts and ranked by the GSFM scores to get top 10 understudied genes for the disease.

3.4 Differentially gene expression analysis of a GEO study

From the many GEO studies with up and down signatures for a disease term from RummaGEO [19], we pick the GEO whose signatures contain most understudied genes found for the disease. We then perform Differentially Gene Expression (DGE) analysis on the gene expression data for the study, GSE63420 for Cholangiocarcinoma. We compute the significantly up and down regulated genes comparing healthy control to Cholangiocarcinoma samples using Limma-voom [16, 15] technique. Significantly expressed genes are determined by p-value <0.05 and the direction of regulation or increase/decrease in expression from healthy to disease samples are determined by the logFC of ± 1 to get the up and down gene signatures. These up and down genes are given as separate inputs to Enrichr [17] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for L2S2 [18] up and down signature search to fetch drug predictions for these differentially expressed genes.

4. Discussion

The present study employed a multi-layered bioinformatic pipeline to uncover genes that are recurrently associated with cholangiocarcinoma (CCA) yet remain poorly characterized in the literature. By intersecting disease-centric gene collections from a broad spectrum of public resources (Enrichr, RummaGEO, Rummagene, Geneshot, MONDO, DO, GWAS Catalog, ClinVar) with quantitative PubMed citation metrics, we identified a set of ten "understudied" genes that are frequently represented in CCA-related gene sets but have comparatively few publications. An independent, model-driven approach using the Gene Set Foundational Model (GSFM) yielded a second, non-overlapping cohort of ten understudied candidates with high predictive scores but low citation counts. Subsequent validation in a publicly available transcriptomic dataset (GSE63420) demonstrated that several of these genes are differentially expressed between healthy biliary epithelium and CCA tissue, providing preliminary functional evidence for their involvement in disease biology.

Biological relevance of the identified genes

The first list of understudied genes (e.g., *MYO5B*, *NHSL3*, *KRT7*, *SOWAHB*, *PTPRF*, *ARHGEF16*, *KRT20*, *MAL2*, *LIAT1*, *SLC44A3*) includes several members of the cytoskeletal and membrane trafficking machinery (*MYO5B*, *MAL2*) as well as keratin isoforms (*KRT7*, *KRT20*) that are known to be expressed in epithelial tissues. Their up-regulation in CCA samples suggests a possible contribution to altered cell polarity, migration, and epithelial-mesenchymal transition—processes that are central to cholangiocarcinogenesis. *SOWAHB* was the sole gene from this set that showed significant down-regulation, hinting at a potential tumor-suppressive role that warrants further investigation.

The GSFM-derived candidates (*CCDC88A*, *RPS6KA3*, *KIF20B*, *RAD54L*, *PTPN14*, *PTPN13*, *EYA4*, *PTPN12*, *WWC1*, *RBBP8*) are enriched for signaling phosphatases (PTPN family), DNA repair factors (RAD54L, RBBP8), and transcriptional regulators (EYA4). Many of these proteins have established roles in other malignancies, yet their specific contributions to CCA have not been

explored. The observation that several phosphatases (PTPN14, PTPN13, PTPN12) are up-regulated aligns with emerging data that dysregulated tyrosine-phosphatase signaling can modulate the tumor microenvironment and immune evasion, both of which are prominent features of CCA.

Integration with pathway enrichment and drug repurposing

Enrichment analysis of the up-regulated gene set highlighted pathways related to cell cycle progression, DNA replication, and focal adhesion, whereas the down-regulated set was enriched for metabolic and bile-acid related processes. These findings are consistent with the known metabolic reprogramming and proliferative drive of CCA cells. The drug-prediction exercise using L2S2 identified several compounds (e.g., erlotinib) that intersect with the identified gene signatures, suggesting that some of the understudied genes may modulate sensitivity to existing targeted therapies. Although the current drug list is limited by the stringent statistical thresholds applied, it provides a starting point for hypothesis-driven pharmacologic screening.

Methodological strengths and limitations

A key strength of this work lies in its systematic combination of literature-based citation filtering with data-driven gene-set augmentation. By leveraging both empirical co-occurrence (gene-set frequency) and predictive modeling (GSFM scores), we mitigated the bias inherent in any single approach. Nonetheless, several limitations must be acknowledged:

- **Citation bias**: PubMed counts reflect research interest rather than biological importance; genes that are newly discovered or studied in non-human systems may be under-counted.
- **Gene-set heterogeneity**: The source databases vary in curation depth and disease specificity, potentially introducing noise into the frequency calculations.
- **Single-cohort validation**: Differential expression was assessed in only one GEO dataset (GSE63420). While this dataset is wellcharacterized, validation across multiple independent cohorts and at the protein level would strengthen the conclusions.
- **Predictive model opacity**: GSFM predictions are derived from large language-model embeddings; the biological rationale for individual scores is not directly interpretable, necessitating experimental confirmation.

Future directions

To translate these computational insights into actionable biology, the following steps are recommended:

1. **Experimental validation**: Perform CRISPR-mediated knockout or siRNA knockdown of the top understudied genes in CCA cell lines and patient-derived organoids to assess effects on proliferation, invasion, and drug response.
2. **Proteomic profiling**: Verify whether transcriptional changes correspond to protein-level alterations using mass-spectrometry or immunohistochemistry on tissue microarrays.
3. **Clinical correlation**: Correlate gene expression with patient outcomes (overall survival, recurrence) in larger, multi-institutional cohorts to evaluate prognostic value.
4. **Mechanistic studies**: Map the signaling networks of phosphatases (PTPN family) and DNA-repair factors (RAD54L, RBBP8) in the context of CCA-specific mutational landscapes (e.g., FGFR2 fusions, IDH1/2 mutations).
5. **Therapeutic exploration**: Test the identified drug candidates, alone or in combination with standard cisplatin–gemcitabine, in pre-clinical models that overexpress the understudied genes.

Conclusion

By integrating literature metrics, gene-set aggregation, and machine-learning-based prediction, we have highlighted a panel of understudied genes that are recurrently implicated in cholangiocarcinoma yet remain largely unexplored. Preliminary expression analyses support their dysregulation in tumor tissue, and pathway enrichment points to biologically plausible roles in cell-cycle control, DNA repair, and signaling modulation. These findings lay the groundwork for targeted functional studies that could uncover novel biomarkers and therapeutic vulnerabilities in this aggressive malignancy.

Acknowledgements

This manuscript used assistance from the Ollama GPT-OSS:120b large language model and DeepDive2.0 resource.

References

[1] Jesus M. Banales, Vincenzo Cardinale, Guido Carpino, and et al. Expert consensus document: Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the european network for the study of cholangiocarcinoma (ens-cca). *Nat Rev Gastroenterol Hepatol*, 2016.

[2] Jesus M. Banales, Jose J. G. Marin, Angela Lamarca, and et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. *Nat Rev Gastroenterol Hepatol*, 2020.

[3] Sumera I. Ilyas, Shahid A. Khan, Christopher L. Hallemeier, and et al. Cholangiocarcinoma - evolving concepts and therapeutic strategies. *Nat Rev Clin Oncol*, 2018.

[4] Donald Maxwell Parkin. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer*, 2006.

[5] Nataliya Razumilava and Gregory J. Gores. Cholangiocarcinoma. *Lancet*, 2014.

[6] Juan Valle, Harpreet Wasan, Daniel H. Palmer, and et al. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *N Engl J Med*, 2010.

[7] Hiromi Nakamura, Yasuhito Arai, Yasushi Totoki, and et al. Genomic spectra of biliary tract cancer. *Nat Genet*, 2015.

[8] Ghassan K. Abou-Alfa, Vaibhav Sahai, Antoine Hollebecque, and et al. Pemigatinib for previously treated, locally advanced or metastatic cholangiocarcinoma: a multicentre, open-label, phase 2 study. *Lancet Oncol*, 2020.

[9] David M. Hyman, Igor Puzanov, Vivek Subbiah, and et al. Vemurafenib in multiple nonmelanoma cancers with braf v600 mutations. *N Engl J Med*, 2015.

[10] Aurelien Marabelle, Dung T. Le, Paolo A. Ascierto, and et al. Efficacy of pembrolizumab in patients with noncolorectal high microsatellite instability/mismatch repair-deficient cancer: Results from the phase ii keynote-158 study. *J Clin Oncol*, 2020.

[11] Eric Tran, Simon Turcotte, Alena Gros, and et al. Cancer immunotherapy based on mutation-specific cd4+ t cells in a patient with epithelial cancer. *Science*, 2014.

[12] Clemens Kratochwil, Paul Flechsig, Thomas Lindner, and et al. 68ga-fapi pet/ct: Tracer uptake in 28 different kinds of cancer. *J Nucl Med*, 2019.

[13] Laura Broutier, Gianmarco Mastrogiovanni, Monique Ma Verstegen, and et al. Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. *Nat Med*, 2017.

[14] A. Lachmann et al. Massive mining of publicly available rna-seq data from human and mouse. *Nature Communications*, 9(1366), 2018.

[15] C. W.and others Law. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15 R29, 2014.

[16] M. E. Ritchie et al. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43, 2015.

[17] Z. Xie. Gene set knowledge discovery with enrichr. *Current Protocols*, 1, 2021.

[18] G. B. Marino et al. L2s2: chemical perturbation and crispr ko lincs l1000 signature search engine. *Nucleic Acids Research*, 53, 2025.

[19] G. B. Marino et al. L2s2: chemical perturbation and crispr ko lincs l1000 signature search engine. *Nucleic Acids Research*, 2025.

[20] D. J. B Clarke et al. Rummagene: massive mining of gene sets from supporting materials of biomedical research publications. *Communications Biology*, 7, 2024.

[21] A. Lachmann et al. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research*, 47, 2019.

[22] N.A.b Vasilevsky et al. Mondo: Integrating disease terminology across communities. *Genetics*, 2025.

[23] L.M. Schriml et al. The human disease ontology 2022 update. *Nucleic Acids Research*, 50, 2022.

[24] A. Buniello et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Research*, 47, 2019.

[25] M.J. Landrum et al. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 2014.

[26] D. J. B. Clarke et al. A gene set foundation model pre-trained on a massive collection of diverse gene sets. *Preprint on bioRxiv*, 2025.