

Under-studied Genes Likely Associated with Cholangiocarcinoma

Abinanda Prabhakaran*

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 Foundation 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 samples) 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 Perturb-Seqr on the combined gene signatures identified several compounds, including nicotinamide, 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, USA.

1. Introduction

Cholangiocarcinoma (CCA) comprises a heterogeneous group of biliary epithelial malignancies that can arise anywhere along the biliary tree and now accounts for roughly 15 % of primary liver cancers and 3 % of gastrointestinal tumours worldwide [1]. Over the past three decades the incidence of intra-hepatic CCA (iCCA) has risen markedly in both high- and low-income regions, a trend that has been documented in population-based registries and attributed to changing risk-factor profiles [1–3]. While classic risk factors such as liver fluke infection, primary sclerosing cholangitis, hepatolithiasis, and exposure to carcinogenic toxins remain important in endemic areas [4–6], a substantial proportion of cases arise in the absence of identifiable aetiologies, suggesting contributions from metabolic liver disease, chronic viral hepatitis, and obesity [6, 7].

Molecular profiling has revealed that CCA is not a single disease entity but rather a collection of subtypes defined by distinct genomic and epigenomic alterations. Recurrent driver events include FGFR2 fusions (predominantly in iCCA), IDH1/2 gain-of-function mutations, KRAS, TP53, BRAF, and ERBB2 amplifications, as well as chromatin-remodelling gene mutations [8–10]. These alterations have informed a new classification scheme that integrates anatomical, genetic, and epigenetic features, thereby guiding precision-medicine approaches [1, 11]. The dense desmoplastic stroma characteristic of CCA further modulates tumour biology, with cancer-associated fibroblasts (CAFs) and fibroblast-activation protein (FAP) contributing to immunosuppression and therapeutic resistance [12, 13]. Therapeutically, systemic chemotherapy with cisplatin plus gemcitabine remains the stan-

standard first-line regimen for advanced biliary tract cancers, including CCA, based on phase III data demonstrating superior progression-free survival compared with gemcitabine alone [14]. The emergence of molecularly targeted agents has begun to improve outcomes for selected patients: FGFR inhibitors such as pemigatinib show activity in FGFR2-rearranged tumours [15], while IDH1 inhibition with ivosidenib prolongs survival in IDH1-mutant disease [16]. Immunotherapy is also gaining traction; pembrolizumab is effective in microsatellite-instability-high/dMMR CCA [17] and, when combined with cisplatin-gemcitabine, improves overall survival in the phase III KEYNOTE-966 trial [18]. These advances underscore the importance of comprehensive genomic testing, for which expert societies now recommend routine next-generation sequencing in advanced CCA [19]. Despite these strides, early diagnosis remains challenging because non-invasive imaging lacks sufficient specificity, and histological confirmation is often required [20, 21]. Emerging diagnostic tools—including fluorescence in-situ hybridisation for aneusomy, organoid-based drug screening platforms, and novel PET tracers such as ^{68}Ga -FAPI—promise to enhance detection and therapeutic selection [22, 23]. Continued integration of molecular insights, refined risk-factor stratification, and multidisciplinary management is essential to improve the dismal prognosis that still characterises cholangiocarcinoma worldwide.

2. Results

After extracting gene sets for Cholangiocarcinoma from various resources including Enrichr, RummaGEO, Rummage, Geneshot, MONDO, DO, GWAS Catalog and ClinVar, we try to identify those genes that are understudied for Cholangiocarcinoma with fewer publications on PubMed. We plot publication counts and gene set counts for each Cholangiocarcinoma gene using only the Cholangiocarcinoma disease gene sets (Figure 1). 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. We plot publication counts and GSFM gene scores for each of the predicted Cholangiocarcinoma genes from GSFM by augmenting the MONDO disease

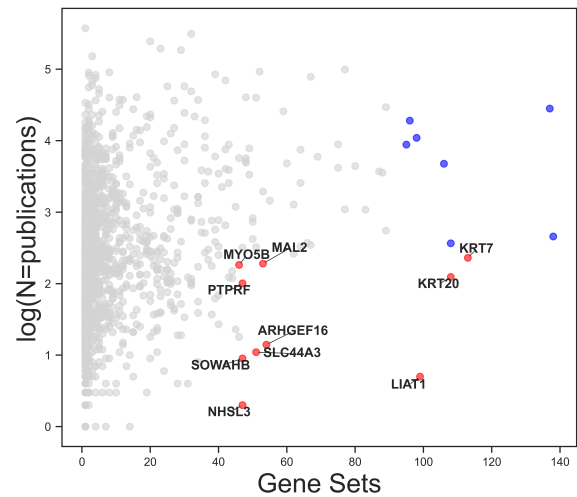


Figure 1. Scatterplot of publication counts vs gene set counts across only Cholangiocarcinoma gene sets for each of the Cholangiocarcinoma genes. Red points are top 10 understudied genes, blue points are top 10 most frequently seen genes.

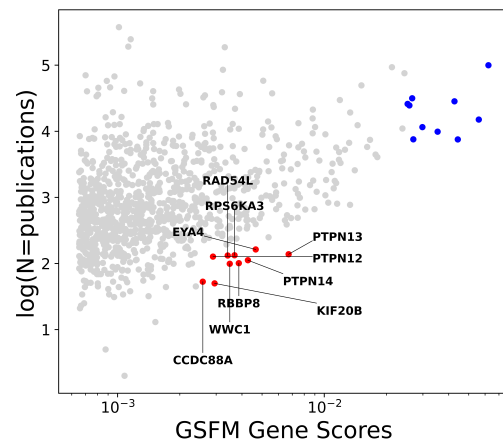


Figure 2. Scatterplot of publication counts vs GSFM gene scores for each of the predicted Cholangiocarcinoma genes. Red points are top 10 understudied genes with high GSFM scores but fewer publications, blue points are top 10 genes with high GSFM scores.

genes for Cholangiocarcinoma (Figure 2). 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 [24] 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 the scatterplot of the first two Principal Components (PCs) of the transformed gene expression data is available for the samples considered for the analysis (Figure 3). To perform DGE analysis, Limma-voom [25, 26] technique is applied 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. A volcano plot shows the DEGs identified for [GSE63420](#) study(Figure 4). 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.

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 [27] to get enriched terms with these DEGs as input queries (Figure 5A, Figure 5B).

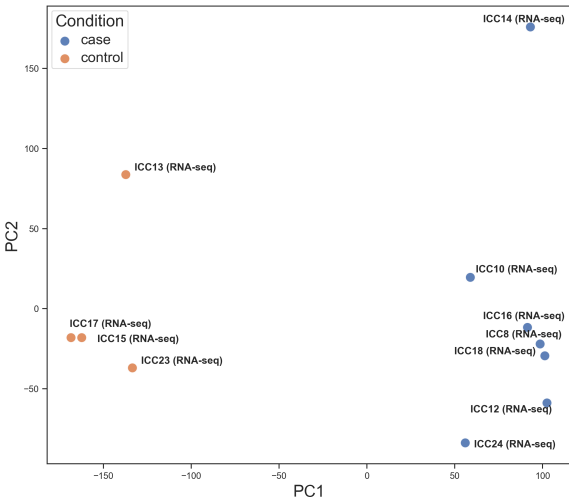


Figure 3. PCA plot of control and disease samples from the GEO study GSE63420. Blue points are control samples and orange points are disease samples. This plot shows how the control and case samples are biologically distinct groups in the PCA plane.

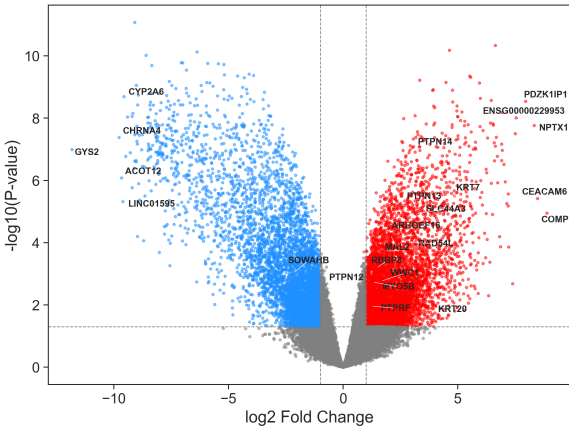


Figure 4. Volcano plot of P-value and LogFC on the limma-voom results for the GEO study for the Healthy Control vs Cholangiocarcinoma samples.

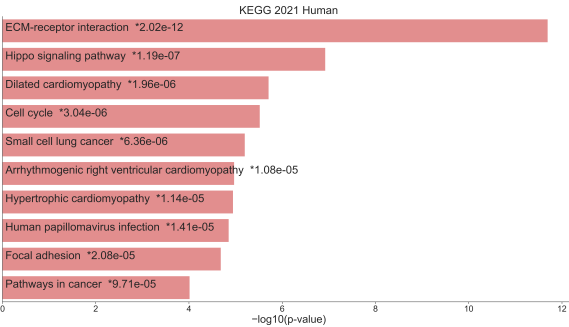


Figure 5. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input up gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input up gene set in the case of Healthy Control vs Cholangiocarcinoma

Using both the up and down genes, we get drugs

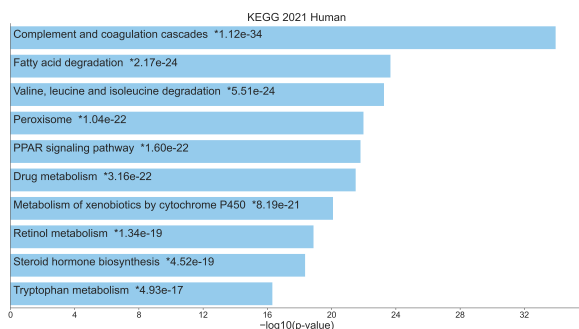


Figure 6. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input down gene set are displayed based on the $-\log_{10}(\text{p-value})$, with the actual p-value shown next to each term. The term at the top has the most significant overlap with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

reversers from Perturb-Seqr [28] associated with the disease gene signatures searched, with details of the predicted drugs (Table 2 and 3).

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.

3.2 Potentially understudied genes from disease-associated genes

The gene sets for the Cholangiocarcinoma disease was extracted from resources - Enrichr [27], RummaGEO [29], Rummagene [30], Geneshot [31], MONDO [32], DO [33], GWAS Catalog [34] and ClinVar [35]. 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 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 Foundation Model (GSFM) [36], to augment the disease genes extracted for the disease from either MONDO [32] or GWAS catalog [34] 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 [29], 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 [25, 26] 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 [27] to fetch enrichment results for the input from KEGG 2021 library and these up and down signatures are given together as input for Perturb-Seqr [28] up and down signature search to fetch drug predictions for these differentially expressed genes.

3.5 Constructing disease reports using the gpt-oss:120b LLM model

Reports are constructed for each disease with abstract, introduction, results, methods, discussion, acknowledgements and references sections. The abstract, introduction and discussion sections are generated by prompting the gpt-oss:120b model, while the other sections are created with templates and custom code. The report starts with an introduction section that

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 KKKU-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 KKKU-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 α 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 α 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

provides a summary with verified citations about the disease using the DeepDive2.0 pipeline. DeepDive2.0 queries the disease term in PubMed, and then uses an LLM to summarize the top 50 most cited articles that mentioned the disease. Followed by the results section

that highlights the understudied genes for each disease found by both the methods discussed above, and the findings from the DGE analysis. A graphical abstract for each report is generated using the Gemini-3.1 flash image model by uploading the entire report to Gemini

Perturbation	Cell line	Dataset	adjPvalueReverse	oddsRatioReverse
tug1		RummaGEO Gene	2.68e-73	3.16
notch3		RummaGEO Gene	2.02e-50	4.95
notch3		RummaGEO Gene	4.55e-40	3.32
notch3		RummaGEO Gene	9.89e-37	3.8
nudt12		RummaGEO Gene	4.53e-34	2.82
notch3		RummaGEO Gene	3.52e-29	3.07
Scd1		CREEDS Gene	2.85e-26	3.74
nudt12		RummaGEO Gene	7.85e-27	3.01
mrs2		RummaGEO Gene	2.59e-23	2.77
EPCAM		CREEDS Gene	2.29e-22	3.31

Table 2. Gene knockout/knockdown reversers from Perturb-Seqr using up and down gene set search

Perturbation	Concentration	Cell line	Dataset	adjPvalueReverse	oddsRatioReverse	approved
nicotinamide			RummaGEO Chem	4.53e-24	2.09	True
nicotinamide			RummaGEO Chem	6.39e-24	2.11	True
cisplatin			RummaGEO Chem	5.29e-21	2.32	True
cisplatin			RummaGEO Chem	1.66e-11	1.97	True
Amsacrine	5.0uM	LOX-IMVI	Tahoe-100M	6.50e-09	4.12	True
dextran			RummaGEO Chem	1.01e-09	3.29	True
estradiol	3.33uM	MCF7	LINCS L1000 CP	1.05e-05	2.55	True
Amsacrine		LS	Tahoe-100M	2.89e-04	3.24	True
atorvastatin	2.5uM	PHH	LINCS L1000 CP	5.12e-04	2.39	True
Mitoxantrone			Tahoe-100M	1.60e-03	3.05	True

Table 3. Drug reversers from Perturb-Seqr using up and down gene set search

manually. So, all of the figures along with the text are then made into a LaTeX bundle to produce the final disease report.

3.6 Generating videos for disease reports using Paper2Video pipeline

To generate videos with generated slides and narration, we used Paper2Video pipeline to automatically convert research publication-like report (given as LaTeX bundle) into a narrated video presentation. For each disease report LaTeX bundle, along with the given reference speaker image, and a short reference audio sample, the system uses LLMs to generate slides, synthesizes per-slide speech via voice cloning (F5-TTS) to produce a final MP4 with subtitles.

3.7 Generating podcasts for disease reports with Gemini-2.5-flash model

To generate podcasts for each disease report, we prompt gemini-2.5-flash model to write an engaging podcast content for entire disease report. Using the default voice available in Edge-TTS, we produce a final audio file MP3 for the generated podcast content.

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, Rummagine, 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 Perturb-Seqr identified several compounds (e.g., nicotinamide and cisplatin) 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] Banales J M, Marin J J G, Lamarca A, et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. *Nature Reviews Gastroenterology & Hepatology*, 17(9), 2020. doi:10.1038/s41575-020-0310-z.
- [2] Shaib Y and El-Serag H. The epidemiology of cholangiocarcinoma. *Seminars in Liver Disease*, 24(02), 2004. doi:10.1055/s-2004-828889.
- [3] Saha S K, Zhu A X, Fuchs C S, et al. Forty-year trends in cholangiocarcinoma incidence in the u.s.: Intrahepatic disease on the rise. *The Oncologist*, 21(5), 2016. doi:10.1634/theoncologist.2015-0446.
- [4] Parkin D M. The global health burden of infection-associated cancers in the year 2002. *International Journal of Cancer*, 118(12), 2006. doi:10.1002/ijc.21731.
- [5] Sripa B, Kaewkes S, Sithithaworn P, et al. Liver fluke induces cholangiocarcinoma. *PLoS Medicine*, 4(7), 2007. doi:10.1371/journal.pmed.0040201.
- [6] Tyson G L and El-Serag H B. Risk factors for cholangiocarcinoma. *Hepatology*, 54(1), 2011. doi:10.1002/hep.24351.
- [7] Michelotti G A, Machado M V, and Diehl A M. Nafld, nash and liver cancer. *Nature Reviews Gastroenterology & Hepatology*, 10(11), 2013. doi:10.1038/nrgastro.2013.183.

- [8] Nakamura H, Arai Y, Totoki Y, et al. Genomic spectra of biliary tract cancer. *Nature Genetics*, 47(9), 2015. doi:[10.1038/ng.3375](https://doi.org/10.1038/ng.3375).
- [9] Borger D R, Tanabe K K, Fan K C, et al. Frequent mutation of isocitrate dehydrogenase(*IDH1*) and *IDH2* in cholangiocarcinoma identified through broad-based tumor genotyping. *The Oncologist*, 17(1), 2011. doi:[10.1634/theoncologist.2011-0386](https://doi.org/10.1634/theoncologist.2011-0386).
- [10] Jusakul A, Cutcutache I, Yong C H, et al. Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma. *Cancer Discovery*, 7(10), 2017. doi:[10.1158/2159-8290.CD-17-0368](https://doi.org/10.1158/2159-8290.CD-17-0368).
- [11] Ilyas S I, Khan S A, Hallemeier C L, et al. Cholangiocarcinoma — evolving concepts and therapeutic strategies. *Nature Reviews Clinical Oncology*, 15(2), 2017. doi:[10.1038/nrclinonc.2017.157](https://doi.org/10.1038/nrclinonc.2017.157).
- [12] Landskron G, De la Fuente M, Thuwajit P, et al. Chronic inflammation and cytokines in the tumor microenvironment. *Journal of Immunology Research*, 2014, 2014. doi:[10.1155/2014/149185](https://doi.org/10.1155/2014/149185).
- [13] Yang X, Lin Y, Shi Y, et al. Fap promotes immunosuppression by cancer-associated fibroblasts in the tumor microenvironment via stat3–ccl2 signaling. *Cancer Research*, 76(14), 2016. doi:[10.1158/0008-5472.CAN-15-2973](https://doi.org/10.1158/0008-5472.CAN-15-2973).
- [14] Valle J, Wasan H, Palmer D H, et al. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *New England Journal of Medicine*, 362(14), 2010. doi:[10.1056/NEJMoa0908721](https://doi.org/10.1056/NEJMoa0908721).
- [15] Abou-Alfa G K, Sahai V, Hollebecque A, et al. Pemigatinib for previously treated, locally advanced or metastatic cholangiocarcinoma: a multicentre, open-label, phase 2 study. *The Lancet Oncology*, 21(5), 2020. doi:[10.1016/S1470-2045\(20\)30109-1](https://doi.org/10.1016/S1470-2045(20)30109-1).
- [16] Abou-Alfa G K, Macarulla T, Javle M M, et al. Ivosidenib in *idh1*-mutant, chemotherapy-refractory cholangiocarcinoma (claridhy): a multicentre, randomised, double-blind, placebo-controlled, phase 3 study. *The Lancet Oncology*, 21(6), 2020. doi:[10.1016/S1470-2045\(20\)30157-1](https://doi.org/10.1016/S1470-2045(20)30157-1).
- [17] Marabelle A, Le D T, Ascierto P A, et al. Efficacy of pembrolizumab in patients with noncolorectal high microsatellite instability/mismatch repair-deficient cancer: Results from the phase ii keynote-158 study. *Journal of Clinical Oncology*, 38(1), 2020. doi:[10.1200/JCO.19.02105](https://doi.org/10.1200/JCO.19.02105).
- [18] Kelley R K, Ueno M, Yoo C, et al. Pembrolizumab in combination with gemcitabine and cisplatin compared with gemcitabine and cisplatin alone for patients with advanced biliary tract cancer (keynote-966): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet*, 401(10391), 2023. doi:[10.1016/S0140-6736\(23\)00727-4](https://doi.org/10.1016/S0140-6736(23)00727-4).
- [19] Mosele F, Remon J, Mateo J, et al. Recommendations for the use of next-generation sequencing (ngs) for patients with metastatic cancers: a report from the esmo precision medicine working group. *Annals of Oncology*, 31(11), 2020. doi:[10.1016/j.annonc.2020.07.014](https://doi.org/10.1016/j.annonc.2020.07.014).
- [20] Razumilava N and Gores G J. Cholangiocarcinoma. *The Lancet*, 383(9935), 2014. doi:[10.1016/S0140-6736\(13\)61903-0](https://doi.org/10.1016/S0140-6736(13)61903-0).
- [21] Ilyas S I and Gores G J. Pathogenesis, diagnosis, and management of cholangiocarcinoma. *Gastroenterology*, 145(6), 2013. doi:[10.1053/j.gastro.2013.10.013](https://doi.org/10.1053/j.gastro.2013.10.013).
- [22] Broutier L, Mastrogiovanni G, Verstegen M M, et al. Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. *Nature Medicine*, 23(12), 2017. doi:[10.1038/nm.4438](https://doi.org/10.1038/nm.4438).
- [23] Kratochwil C, Flechsig P, Lindner T, et al. ⁶⁸ga-fapi pet/ct: Tracer uptake in 28 different kinds of cancer. *Journal of Nuclear Medicine*, 60(6), 2019. doi:[10.2967/jnumed.119.227967](https://doi.org/10.2967/jnumed.119.227967).
- [24] Lachmann A et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, 9(1), 2018. doi:[10.1038/s41467-018-03751-6](https://doi.org/10.1038/s41467-018-03751-6). URL <https://archs4.org/>.
- [25] Law CW et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 2014. doi:[10.1186/gb-2014-15-2-r29](https://doi.org/10.1186/gb-2014-15-2-r29).
- [26] Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. doi:[10.1093/nar/gkv007](https://doi.org/10.1093/nar/gkv007).
- [27] Xie Z et al. Gene set knowledge discovery with Enrichr. *Current Protocols*, 1(3), 2021. doi:[10.1002/cpz1.90](https://doi.org/10.1002/cpz1.90). URL <https://maayanlab.cloud/Enrichr/>.
- [28] Gardner JK et al. Perturb-Seqr [internet]. <https://perturbseqr.maayanlab.cloud/>, 2026.

- [29] Marino GB et al. RummaGEO: Automatic mining of human and mouse gene sets from GEO. *Patterns*, 5(10), 2024. doi:10.1016/j.patter.2024.101072. URL <https://rummageo.com/>.
- [30] Clarke DJB et al. Rummagene: massive mining of gene sets from supporting materials of biomedical research publications. *Communications Biology*, 7(1), 2024. doi:10.1038/s42003-024-06177-7. URL <https://rummagene.com/>.
- [31] Lachmann A et al. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research*, 47(W1), 2019. doi:10.1093/nar/gkz393. URL <https://maayanlab.cloud/geneshot/>.
- [32] Vasilevsky NA et al. Mondo: Integrating disease terminology across communities. *Genetics*, 232(4), 2025. doi:10.1093/genetics/iyaf215. URL <https://mondo.monarchinitiative.org/>.
- [33] Schriml LM et al. The human disease ontology 2022 update. *Nucleic Acids Research*, 50(D1), 2021. doi:10.1093/nar/gkab1063. URL <https://www.disease-ontology.org/>.
- [34] Sollis E et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), 2022. doi:10.1093/nar/gkac1010. URL <https://www.ebi.ac.uk/gwas/>.
- [35] Landrum MJ et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1), 2013. doi:10.1093/nar/gkt1113. URL <https://www.ncbi.nlm.nih.gov/clinvar/>.
- [36] Clarke DJB et al. A gene set foundation model pre-trained on a massive collection of diverse gene sets. *Preprint on bioRxiv*, 2025. doi:10.1101/2025.05.30.657124. URL <https://gsfm.maayanlab.cloud/>.