

Liver Cancer Screening Tool Using Gut Microbe Composition

Yonathan Wagaye
Computer Science, NYUAD
yaw2002@nyu.edu

Advised by: Mai Oudah, Mohamad Eid

ABSTRACT

Background. Liver cancer is the world's third-leading cause of cancer mortality, with more than 80 % of cases arising in East Asia and sub-Saharan Africa [8]. Earlier, non-invasive biomarkers for hepatocellular carcinoma (HCC) are urgently needed.

Methods. We profiled the gut microbiota by 16S rRNA sequencing in two independent cohorts—120 Thai adults (76 controls, 25 HCC, 19 iCCA) and 57 Chinese adults (19 controls, 38 HCC)—using closed-reference OTU picking. From these we built three balanced case-control datasets:

- Thai cohort: 120 adults (76 controls, 25 HCC, 19 iCCA)
- Chinese cohort: 57 adults (19 controls, 38 HCC)

Construction of three balanced case-control datasets:

- **Thai LC:** 44 liver cancer (25 HCC + 19 iCCA) vs. 44 controls (random subset of Thai controls)
- **Thai HCC:** 25 Thai HCC vs. 25 Thai controls
- **Thai + Chinese HCC:** 63 HCC (25 Thai + 38 Chinese) vs. 63 Thai controls

Optional augmentation of taxonomic features with PICRUSt2-inferred MetaCyc pathway abundances. Classification pipeline embedding information-gain and correlation filters inside five classifiers (Random Forest, Naïve Bayes, SVM, logistic regression, J48) under stratified 5-fold cross-validation

Results. Thai cohort. Naïve Bayes + information gain reached 90 % accuracy (AUC = 0.88).

Merged Thai + Chinese. Naïve Bayes (no feature selection)

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك أبوظبي

 NYU | ABU DHABI

Capstone Seminar, Spring 2025, Abu Dhabi, UAE

© 2025 New York University Abu Dhabi.

and Random Forest + information gain tied at 80.8% accuracy.

Merged + pathway layer. Adding MetaCyc pathways and correlation filtering raised accuracy to 84.6% (Random Forest, AUC = 0.83).

Key HCC biomarkers were enriched *Veillonellaceae* and other Proteobacteria-associated genera, with depletion of butyrate-producing *Lachnospiraceae* and *Ruminococcaceae*. Functionally, HCC stool showed higher acetyl-CoA / mevalonate and menaquinone biosynthesis, while controls showed elevated methanogenesis from acetate and sulfide detoxification pathways.

Conclusion. A hybrid taxon + pathway model generalises across populations and offers an accurate, non-invasive HCC screening tool that mechanistically links gut dysbiosis to hepatocarcinogenesis.

KEYWORDS

Machine Learning, Meta-genomics Analysis, Hepatocellular Carcinoma, Intrahepatic cholangiocarcinoma, Gut microbes

Reference Format:

Yonathan Wagaye. 2025. Liver Cancer Screening Tool Using Gut Microbe Composition. In *NYUAD Capstone Seminar Reports, Spring 2025, Abu Dhabi, UAE*. 9 pages.

1 INTRODUCTION

Liver cancer remains a major global health challenge, ranking as the sixth most frequently diagnosed malignancy and the third leading cause of cancer-related death[8]. Among primary liver tumours, hepatocellular carcinoma (HCC) is by far the dominant histological subtype, accounting for roughly 80% of cases; intra-hepatic cholangiocarcinoma (iCCA) and other rarer entities comprise the remainder[15]. In 2018 an estimated 661 000 new HCC cases were reported worldwide, with the highest incidence in Eastern and South-Eastern Asia and parts of sub-Saharan Africa, where chronic hepatitis B/C, aflatoxin exposure, alcohol misuse, and metabolic disease are endemic risk factors [15]. Despite advances in antiviral

therapy and imaging, current diagnosis still hinges on serum alpha-fetoprotein plus ultrasound or MRI—costly, operator dependent modalities that often detect tumors only at intermediate stages. Hence, non-invasive, scalable biomarkers that enable earlier HCC detection are urgently needed.

1.1 Gut microbiome as a biomarker source

The gut–liver axis offers a compelling window into hepatic physiology: portal flow constantly delivers microbial metabolites and inflammatory signals to hepatocytes. Multiple studies report that HCC is accompanied by gut dysbiosis—typically an expansion of lipopolysaccharide-rich Proteobacteria and depletion of short chain fatty acid producing Firmicutes [13, 14]. Recent reviews underscore the promise of faecal microbial signatures as non-invasive predictors of liver and pancreatic cancers [7], while new sequencing technologies and computational pipelines enable increasingly rich functional profiling of microbial communities [6].

1.2 Machine-learning opportunity and challenges

Machine learning (ML) can extract predictive structure from the high-dimensional, compositional data produced by 16S rRNA or metagenomic sequencing. Classifiers such as Random Forest, Support-Vector Machines, Naïve Bayes, and Logistic Regression—especially when paired with feature-selection filters like information gain or correlation-based subset selection—have delivered robust diagnostic models for other microbiome associated diseases [10, 11]. Translating these gains to HCC screening, however, requires addressing two key challenges:

- (1) **Inter-individual variability:** Microbiome composition shifts with diet, geography, and host genetics, complicating the development of consensus biomarkers [19].
- (2) **High dimensionality:** OTU or amplicon sequence variable tables often contain tens of thousands of features, predisposing naive models to overfitting.

1.3 Study objectives

This work focuses exclusively on HCC and pursues three objectives:

- (1) Develop a gut-microbiome-based ML classifier that discriminates HCC from healthy controls using stool samples.
- (2) Assess cross-regional generalisability by merging Thai and Chinese cohorts.
- (3) Test whether augmenting taxonomic features with PICRUSt2-inferred MetaCyc pathways (hybrid taxon + pathway models) outperforms taxonomy-only models.

2 RELATED WORK

Early work demonstrated that gut dysbiosis accompanies liver disease and may yield actionable biomarkers for hepatocellular carcinoma (HCC). Kaźmierczak-Siedlecka et al. [7] reviewed multiple studies linking microbial metabolites—such as lipopolysaccharide (LPS), secondary bile acids, and short-chain fatty acids (SCFAs)—to hepatic inflammation and tumor promotion, highlighting pattern-recognition receptors like TLR4 as putative diagnostic targets.

In one of the largest 16S-based surveys, Ren et al. [14] applied a Random Forest classifier to 486 Chinese stool samples and identified 30 genus-level taxa (e.g., blooms of *Fusobacterium* and depletion of *Faecalibacterium*) that separated HCC from controls with AUC 0.92, though species-level resolution remained limited. Complementing this, Pomyen et al. [13] used whole-genome shotgun sequencing on 120 Thai samples (HCC, iCCA, controls) and linear discriminant analysis (LDA) to nominate species such as *Veillonella atypica*, *Ruminococcus gnavus*, and *Adlercreutzia* as HCC markers, and linked bile-acid–transforming enzymes (e.g., 7-dehydroxylase) to carcinogenesis, but did not develop a predictive ML model. Machine-learning benchmarks have further refined these insights. Zhang et al. [19] predicted micro-vascular invasion in HCC using Random Forest (AUC = 0.95) but noted geographic microbiome drift as a barrier to model transferability.

Beyond taxonomy, functional and metabolomic studies have shown enzymatic drivers of HCC. Luo et al. [8] and Guertin & Wellen [5] reviewed microbial enzymes—such as acetyl-CoA carboxylase for lipid anabolism and bile-salt hydrolases—that rewire host metabolism toward tumorigenesis. Feng et al. [4] integrated untargeted LC-MS metabolomics with 16S in a 57-sample Chinese cohort, mapping correlations among stool metabolites (e.g., retinol derivatives), genera (*Streptococcus*, *Veillonella*), and serum -fetoprotein; they found enrichment of steroid-hormone and retinol-metabolism pathways in HCC but did not build a standalone microbiome classifier.

Collectively, these studies (i) confirm robust taxonomic and functional HCC signatures, (ii) achieve high accuracy within individual cohorts, yet (iii) lack cross-regional validation and integrated pathway-aware classifiers. The present work addresses these gaps by merging Thai and Chinese 16S datasets, focusing exclusively on HCC, and appending PICRUSt2-inferred MetaCyc pathways to taxonomic features, thereby delivering a cost-effective, geographically robust, and mechanistically interpretable screening pipeline.

3 METHODOLOGY

Our workflow converts raw sequencing reads from two independent cohorts into four analytically distinct datasets—(a)

Thai LC, (b) Thai HCC, (c) Thai + Chinese, and (d) Thai + Chinese + Pathway—then benchmarks several supervised classifiers under nested cross-validation.

3.1 Dataset assembly

- **a. Thai LC:**
 - Thai cohort (Pomyen et al. [13])
 - HCC + iCCA samples merged into single “liver-cancer” label
 - 44 HCC / 44 controls
- **b. Thai HCC:**
 - Same Thai cohort, with iCCA samples removed
 - HCC-specific analysis
 - 25 HCC / 25 controls
- **c. Thai + Chinese:**
 - Thai HCC (25) + Chinese HCC (38) + Thai controls (63) [4]
 - Balanced 63 HCC / 63 controls
 - Cross-regional testing
- **d. Thai + Chinese + Pathway:**
 - Dataset (c) augmented with PICRUST2-inferred Meta-Cyc pathways
 - Same sample counts (63 HCC / 63 controls)
 - Functional context for classification

3.2 Sequence processing and OTU generation

- (1) **Sequence retrieval & conversion**
For the Thai SRA archives we applied FASTQ (SRA Toolkit v3.1.1) to extract the 16S rRNA samples from the whole metagenome; Chinese samples were already in 16S rRNA format.
- (2) **Quality trimming.**
Trimmomatic v0.40 (SLIDINGWINDOW:4:30, MINLEN:80).
- (3) **16S extraction (Thai only)**
SortMeRNA v2.1 against Greengenes 13_8; top hit per read kept.
- (4) **Closed-reference OTU picking**
QIIME2 v2024.5, vsearch cluster features closed reference at 97%
- (5) **Normalisation**
Convert counts to relative abundance and multiply each abundance by for scaling 100 000.

3.3 Functional-pathway inference

For datasets (b) and (c) we run PICRUST2 v2.6.4 on raw OTU counts and obtained activated pathways for each sample. Then, pathway counts were rescaled to relative abundance and multiplied by 100 000. Finally, we concatenated pathways with OTUs as features in dataset (d).

3.4 Feature selection

Two filter methods applied inside each training fold:

- **Information Gain (IG):** InfoGainAttributeEval + Ranker, threshold 0.03.
- **Correlation-based (CFS):** CfsSubsetEval + BestFirst, backward search, merit 0.85.

3.5 Classification models & settings

We benchmark five classifiers in WEKA 3.9 under nested 5-fold and 10- fold cross validation):

- **Naïve Bayes**
- **Random Forest**
- **Support Vector Machine**
- **Logistic Regression**
- **J48 decision tree**

For dataset (a) we used 10-fold CV; for (b), (c), (d) we applied 5-fold nested CV.

3.6 Evaluation metrics

We report:

- **Accuracy** and **macro-F1** (balanced classes).
- **AUC** (area under ROC).

3.7 Software environment

- Python 3.12, scikit-learn 1.5, pandas 2.2
- R 4.4 with ALDEx2 1.32
- WEKA 3.9.10 (CLI)
- QIIME2 2024.5, PICRUST2 2.6.4, Trimmomatic 0.40, SortMeRNA 2.1, SRA Toolkit 3.1.1

Dataset: Thai LC (44 HCC / 44 CTR)		
Attribute	Information Gain	Correlation-Based
Initial Number of Attributes	23,585	23,585
Avg. Selected Number of Attributes	687	967
Percentage Reduction (%)	97.09	95.90

Table 1: Comparison of Feature Selection Methods: Information Gain vs. Correlation-Based.

4 EXPERIMENTAL ANALYSIS

4.1 Experimental Settings

We evaluated all four datasets using the WEKA 3.9 machine-learning framework, applying the same preprocessing and classification pipeline described in Section 3. For each dataset

Dataset: Thai HCC (25 HCC / 25 CTR)		
Attribute	Information Gain	Correlation-Based
Initial Number of Attributes	23,586	23,586
Avg. Selected Number of Attributes	336	2067
Percentage Reduction (%)	98.6	91.2

Table 2: Feature Selection Summary for Thai HCC Dataset.

Dataset: Thai + Chinese (63 HCC / 63 CTR)		
Attribute	Information Gain	Correlation-Based
Initial Number of Attributes	26,655	26,655
Avg. Selected Number of Attributes	3287	2152
Percentage Reduction (%)	87.7	91.9

Table 3: Feature Selection Summary for Thai + Chinese Dataset.

Dataset: Thai + Chinese + Pathway (63 HCC / 63 CTR)		
Attribute	Information Gain	Correlation-Based
Initial Number of Attributes	26,675	26,675
Avg. Selected Number of Attributes	2171	3304
Percentage Reduction (%)	91.8	87.61

Table 4: Feature Selection Summary for Thai + Chinese + Pathway Dataset.

we isolated a random 20% hold-out set for final testing, and used cross-validation on the remaining 80% for feature selection and model training. In all cases we compared three feature-selection conditions: no selection (baseline), Information Gain (IG), and Correlation-Based Feature Selection (CFS).

a. Thai LC (44 HCC / 44 CTR). Data from Pomyen [13] were processed with SRA Toolkit, Trimmomatic, SortMeRNA, QI-ME2 OTU-picking and normalization as in Section 3. We

applied 10-fold cross-validation to benchmark seven classifiers (Naïve Bayes, Random Forest, SMO, J48, Logistic Regression) under each feature-selection method, and reported mean AUC and accuracy across folds.

b. Thai HCC (25 HCC / 25 CTR). Using the same Thai cohort with iCCA samples removed, we retained 23 586 OTU features and evaluated five classifiers (Naïve Bayes, Random Forest, SMO, Logistic Regression, J48) in WEKA. Feature selection (IG vs CFS) was performed per fold. Models were trained on the selected features and validated on the held-out partition.

c. Thai + Chinese (63 HCC / 63 CTR). We merged the Thai HCC samples with 38 Chinese HCC and 63 Thai controls [4], yielding 126 samples. No 16S extraction was needed for the Chinese FASTQs; all samples underwent closed-reference OTU picking and normalization. The same nested 5-fold CV procedure (inner for IG/CFS, outer for evaluation) was applied to the five classifiers, allowing us to assess cross-regional generalisability.

d. Thai + Chinese + Pathway (63 HCC / 63 CTR). Starting from dataset (c), we ran PICRUSt2 v2.6.4 on the raw OTU counts to infer 467 MetaCyc pathway features and appended these to the taxonomic matrix. The combined 26 675 features were then subjected to the identical nested 5-fold CV and feature-selection workflow in WEKA, enabling direct comparison of taxonomy-only vs hybrid taxon + pathway models.

4.2 Experiment Results

When the positive class comprised both hepatocellular carcinoma (HCC) and intra-hepatic cholangiocarcinoma (iCCA) in the Thai LC set, the strongest configuration—Logistic Regression with correlation-based feature selection—reached **78 % accuracy** and an AUC of 0.81. After removing the iCCA samples (Thai HCC set) label purity produced an immediate lift: Naïve Bayes with Information Gain attained **90 % accuracy** and an AUC of 0.88. The same feature-selection filters that had previously harmed performance now helped, confirming that iCCA and HCC are driven by partly distinct microbial profiles and that mixing them confuses the model.

Merging the Chinese cohort increased the sample size to 126 and expanded the OTU vocabulary from roughly 23,600 to **26,000** features. Baseline Naïve Bayes already performed well (80.8 % accuracy, AUC 0.81), and Information-Gain or CFS offered only marginal gains—an indication that diet- and geography-specific taxa add noise that simple filters cannot fully reconcile. Nevertheless, collapsing the feature space to about half more than halved training time while preserving AUC.

Model	Baseline			Information Gain			Correlation-Based		
	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1
J48	56.81	0.5656	0.5577	63.89	0.6225	0.6288	60.83	0.5925	0.5982
Logistic Regression	75.97	0.7748	0.7495	70.69	0.7193	0.7022	78.06	0.8113	0.7742
Naïve Bayes	78.33	0.7722	0.7747	69.31	0.7302	0.6782	69.17	0.7503	0.6778
SMO	77.50	0.7800	0.7543	77.50	0.7800	0.7543	56.94	0.5675	N/A
Random Forest	70.00	0.78	35.39	60.00	0.60	36.90	70.00	0.80	30.14

Table 5: Performance metrics on the Thai LC dataset.

Model	Baseline			Information Gain			Correlation-Based		
	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1
J48	70.00	0.70	36.15	70.00	0.70	34.94	70.00	0.70	34.94
Logistic Regression	50.00	0.58	40.18	70.00	0.78	44.06	60.00	0.59	34.06
Naïve Bayes	60.00	0.70	37.25	90.00	0.88	46.25	60.00	0.60	35.81
Random Forest	70.00	0.78	35.39	60.00	0.60	36.90	70.00	0.80	30.14
SMO	50.00	0.50	0.36	70.00	0.70	0.52	70.00	0.70	0.38

Table 6: Performance metrics on the Thai HCC dataset.

Model	Baseline			Information Gain			Correlation-Based		
	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1
J48	76.92	0.77	67.07	73.08	0.73	70.67	76.92	0.80	68.90
Logistic Regression	69.23	0.61	69.52	50.00	0.44	60.34	57.69	0.59	60.37
Naïve Bayes	80.77	0.81	77.38	73.08	0.73	71.52	76.92	0.77	72.29
Random Forest	76.92	0.91	72.79	80.77	0.89	72.70	76.92	0.89	73.05
SMO	73.08	0.73	0.70	69.23	0.69	0.68	69.23	0.69	0.71

Table 7: Performance metrics on the Thai + Chinese dataset.

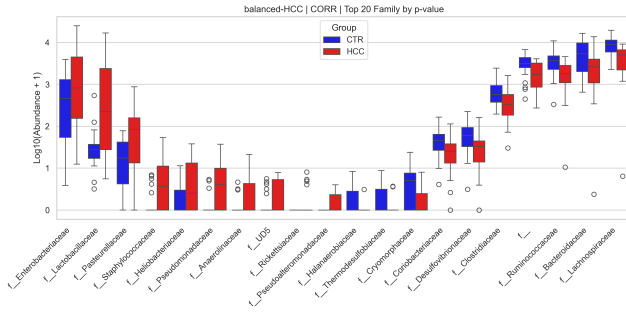
Model	Baseline			Information Gain			Correlation-Based		
	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1
J48	76.92	0.77	67.07	73.08	0.73	70.67	76.92	0.80	68.90
Logistic Regression	69.23	0.61	66.32	50.00	0.43	61.05	57.69	0.57	57.59
Naïve Bayes	80.77	0.81	77.38	73.08	0.75	71.52	76.92	0.77	72.29
Random Forest	73.08	0.89	70.30	80.77	0.86	73.71	84.62	0.87	74.35
SMO	73.08	0.73	0.70	69.23	0.69	0.68	69.23	0.69	0.71

Table 8: Performance metrics on the Thai + Chinese + Pathway dataset.

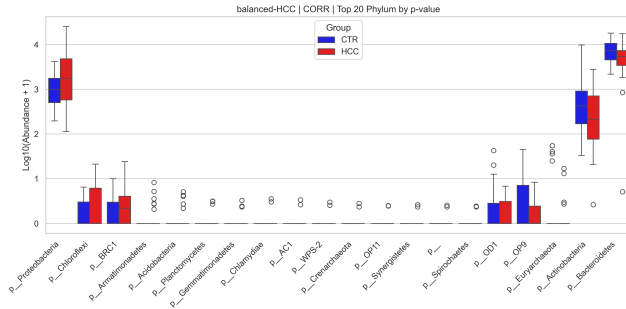
Adding functional context recovered the lost ground. When 467 PICRUSt2 MetaCyc pathways were concatenated with the Thai–Chinese taxonomy, the best Random Forest + CFS model rose to **84.6 % accuracy** with an AUC of 0.87. Pathways linked to acetyl-CoA/mevalonate and menaquinone biosynthesis—previously implicated in tumour lipogenesis

and redox control [5, 8]—appeared among the top-ranked features, suggesting that functional signals stabilise prediction across regions even when taxa vary.

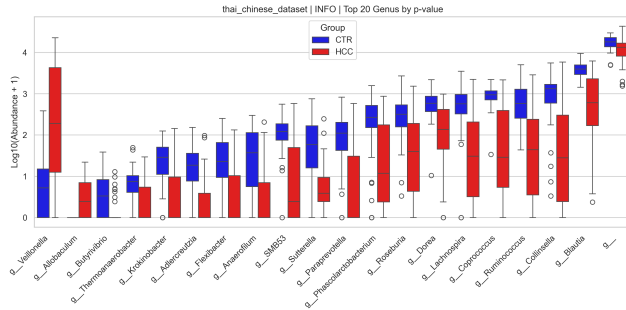
In summary, (i) removing subtype noise yields the largest single gain (78 % to 90 %), (ii) cross-regional integration maintains respectable accuracy but reduces the marginal value of filter methods, and (iii) pathway augmentation restores—and in some cases surpasses—the predictive power



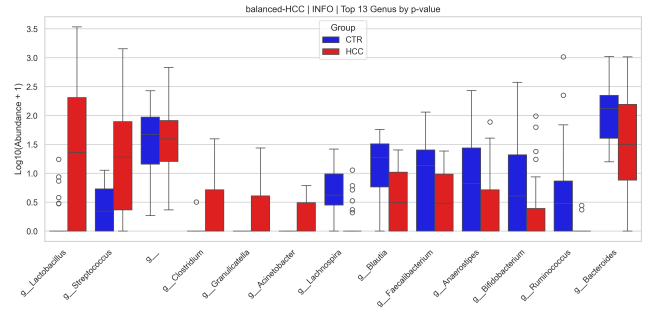
Thai HCC Family - Correlation Based



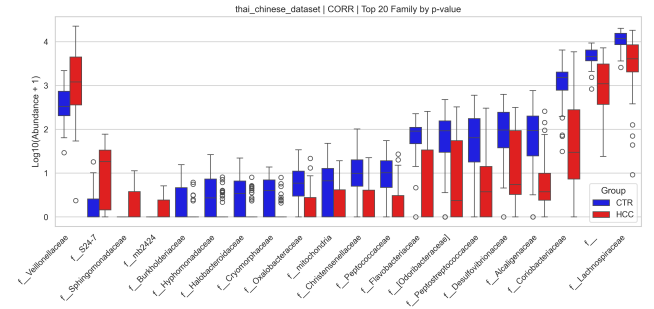
Thai HCC Phylum - Correlation Based



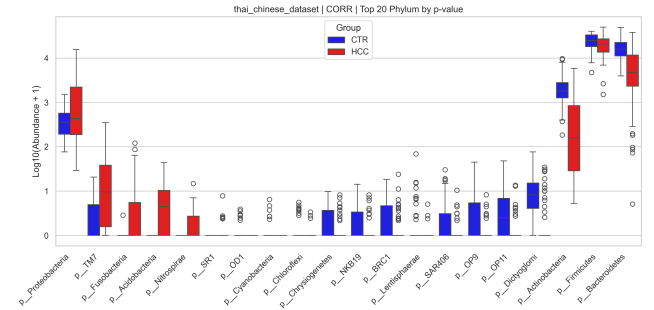
Thai-Chinese HCC Genus - Information Gain



Thai HCC Genus - Information Gain



Thai-Chinese HCC Family - Correlation Based



Thai-Chinese HCC Phylum - Correlation Based

Figure 1: Relative abundance analysis at Phylum, Family, and Genus levels. The plots compare the Information Gain and Correlation-Based feature selection methods across different taxonomic levels.

of single-cohort models while providing mechanistic insight into HCC-associated dysbiosis.

5 DISCUSSION

In the original Thai LC cohort, where HCC and iCCA were pooled, the gut microbiome community already exhibited a carcinogenic tilt. Members of *Proteobacteria* and *Fusobacteria* expanded, while health-associated *Actinobacteria* contracted. Within this pro-inflammatory backdrop, families such as

Enterococcaceae and *Actinomycetaceae* showed higher abundance in cancer stools, similar to reports in cirrhosis and biliary malignancy [1]. Conversely, the depletion of *Prevotellaceae* matched earlier observations that anti-inflammatory fibre fermenters fade as liver pathology advances. At genus rank, the Thai shotgun study by Pomyen [13] is recapitulated: HCC/iCCA samples are rich in *Veillonella* and poor in *Ruminococcus*. A novel finding is the prominence of *Adlercreutzia* in controls; this polyphenol-converting genus has

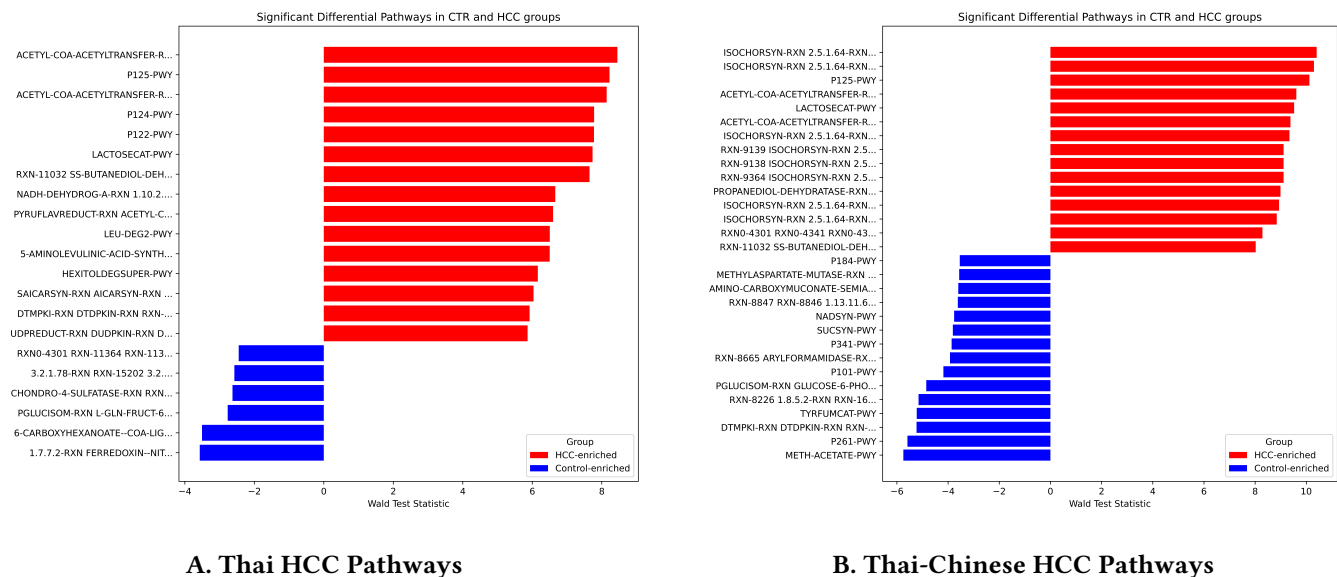


Figure 2: Tornado plots of significant ($\text{padj} < 0.05$) microbial pathways in HCC vs. control for the Thai (A) and Thai-Chinese (B) cohorts; red bars indicate HCC-enriched functions, blue bars control-enriched.

anti-inflammatory credentials in metabolic disease [12] but has not previously been linked to liver-cancer protection.

When the label is purified to HCC only (Thai HCC dataset), the same phylum-level pattern intensifies. *Proteobacteria* now dominate cases, propelled by surges in *Enterobacteriaceae*, *Pasteurellaceae* and lactate-loving *Veillonellaceae* [1, 14]. Genera such as *Veillonella*, *Lactobacillus* and opportunistic *Pantoea* mark the tumour signature, while controls consolidate a butyrate-rich core of *Ruminococcus*, *Blautia* and *Faecalibacterium*. The sharper separation supports growing evidence that iCCA and HCC harbour distinct microbiome drivers and that LPS-heavy *Proteobacteria* foster the chronic inflammation feeding hepatocarcinogenesis [8].

Adding 38 Chinese HCC stools (Thai + Chinese dataset) raises diversity yet preserves many Thai signals. *Veillonella* remains the most robust HCC genus, accompanied by lesser-known but case-enriched *Allobaculum* and *Porphyromonas*. Protective taxa are again dominated by short-chain fatty-acid producers—*Blautia*, *Lachnospira*, *Ruminococcus*. At phylum level, the HCC stool still carries an overgrowth of *Proteobacteria* and *Fusobacteria*, whereas *Firmicutes* and *Bacteroidetes* centre the healthy gut. The persistence of these microbiomes despite dietary and regional heterogeneity underscores their biological sturdiness and echoes multi-centre Chinese findings by Ren et al.[14].

Across all datasets a consensus therefore emerges. **Positive (HCC) biomarkers** include endotoxin-rich *Proteobacteria*, especially *Enterobacteriaceae*, the lactate-metabolising

genus *Veillonella*, and hydrogen-sulphide-producing *Fusobacterium*. These organisms supply LPS or genotoxic metabolites that damage hepatic tissue and amplify Toll-like-receptor cascades. **Negative (protective) biomarkers** cluster among butyrate and acetate producers in *Firmicutes* and *Bacteroidetes*: *Ruminococcus*, *Blautia*, *Lachnospira*, *Bacteroides*. Two genera—control-enriched *Adlercreutzia* and HCC-associated *Allobaculum*—have not been mention in liver-tumour literature before and thus surface here as *novel* findings. Taken together, the taxonomic data paint a coherent ecological picture: HCC is accompanied by an inflammatory shift towards LPS, ethanol and sulphide producers and away from fibre-fermenting, barrier-protective commensals, a pattern that holds even after broadening the cohort geography.

The metabolic pathway construction strengthens the ecological picture sketched by our compositional data. In the Thai-HCC cohort, the most pronounced case-enriched functions are the convergent acetyl-CoA/mevalonate modules (MetaCyc PWY-922, PWY 5910), together with 2,3 butanediol ethanol fermentation and branched chain amino acid catabolism. These routes channel microbial carbon and reducing power into lipogenesis and histone acetylation—processes that accelerate hepatocyte proliferation and tumour epigenetic remodelling [2, 3, 5]. The same stools are taxonomically dominated by *Veillonellaceae* and other *Proteobacteria* that thrive on lactate and deliver lipopolysaccharide; thus functional and compositional layers converge on a pro-inflammatory, lipid-fuelled niche that has been reported in multiple Asian HCC cohorts [14]. Controls, by contrast,

over-express biotin biosynthesis and UDP-GlcNAc production—pathways that support fatty-acid oxidation and epithelial glycosylation and align with their enrichment in butyrate-producing Ruminococcaceae and Blautia [9].

When Chinese cases are combined with the Thai dataset, we continue to see strong enrichment of acetyl-CoA metabolism in HCC patients—a known marker of tumor activity. Furthermore, new cancer related pathways also become apparent, most notably menaquinone (vitamin K) biosynthesis (PWY-5837). Although vitamin K₂ (especially its MK-4 isoform) is widely reported to inhibit tumour growth, promote apoptosis and suppress NF- κ B-mediated inflammation, our pathway analysis reveals that MK-4 biosynthetic potential is *higher* in HCC stools [17, 18]. This apparent paradox is best explained by a shift in community composition: HCC patients show a marked expansion of *Enterobacteriaceae*, *Lactobacillus* and *Bacillus* species—taxa that rely on menaquinones as electron carriers and therefore up-regulate the complete menA-menG pathway [16]. Hence the elevated MK-4 signal more likely reflects an increased *microbial supply* of vitamin K₂ rather than a protective host response, a view supported by recent evidence that gut bacteria capable of synthesising long-chain vitamin K₂ homologues can modulate cancer risk and progression.

Taken together, the data portray a coherent gut–liver axis reproducible across geography. HCC is accompanied by a bloom of LPS-rich Proteobacteria and Fusobacteria whose metabolic outputs—acetate, mevalonate, MK-4, ethanol and peptidoglycan fragments—stimulate lipogenesis, histone acetylation and innate-immune signalling. Conversely, healthy controls maintain a butyrate-centred, methanogen-buffered ecosystem that supports epithelial integrity and redox balance. The recurrence of these pathway–taxon couplings in both single-centre and multi-regional cohorts suggests they are core features of HCC-associated dysbiosis rather than artefacts of local diet, and highlights acetyl-CoA and menaquinone metabolism as tractable microbe-derived targets for early intervention.

6 CONCLUSION

This study leveraged two publicly available gut-microbiome cohorts to build four progressively richer datasets—Thai LC, Thai-HCC, Thai + Chinese HCC, and Thai + Chinese HCC augmented with predicted MetaCyc pathways—and benchmarked a uniform, nested-CV machine-learning pipeline. Purifying the positive label from “liver cancer” to HCC alone raised performance from 78% (AUC 0.81) to 90% (AUC 0.88), while supplementing the multi-regional dataset with functional features restored accuracy to 84% and maintained an AUC of 0.87.

Across every taxonomic resolution a coherent signature emerged: HCC stools are enriched in endotoxin-rich *Proteobacteria* (notably *Enterobacteriaceae*), lactate-utilising *Veillonella* and sulphide-producing *Fusobacterium*, whereas controls harbour butyrate- and acetate-generating *Ruminococcus*, *Blautia*, *Lachnospira* and *Bacteroides*. Two genera surface as novel signals—control-associated *Adlercreutzia* and tumour-associated *Allobaculum*. Functionally, both HCC groups display a dominant acetyl-CoA/mevalonate programme that feeds lipogenesis and epigenetic acetylation, while the Thai + Chinese cases additionally over-produce menaquinone (vitamin K₂). Control samples consistently up-regulate biotin biosynthesis and UDP-GlcNAc production, pathways that support fatty-acid oxidation and mucosal glycosylation.

Together these results portray an HCC-linked gut ecology characterised by LPS, ethanol and sulphide output at the expense of fibre-fermenting, barrier-protective commensals—a pattern robust to geography and further confirmed by pathway analysis. The taxa and the recurrent acetyl-CoA/mevalonate and menaquinone modules highlight tractable, microbe-derived targets for non-invasive screening and for adjunctive intervention in hepatocellular carcinoma.

REFERENCES

- [1] Naga S. Betrapally, Patrick M. Gillevet, and Jasmohan S. Bajaj. 2016. Gut Microbiome and Liver Disease. *Clinical Liver Disease* 8, 4 (2016), 91–94. <https://doi.org/10.1002/cld.452>
- [2] Panpan Chen, Maria Torralba, Jun Tan, Mallory Embree, Karsten Zengler, Catherine Lozupone, and Bernd Schnabl. 2020. Intestinal microbiota-produced ethanol worsens liver injury in mice fed a high-fat diet. *Cell Metabolism* 31, 5 (2020), 941–955.e7. <https://doi.org/10.1016/j.cmet.2020.03.006>
- [3] Renata E. Ericksen, Hyojin Lim, Hyun Woo Kim, Yongjun Shin, Jin Man Kim, Kimoon Lee, Yong-Il Moon, and Jeong Hoon Kim. 2019. Loss of BCAA catabolism during carcinogenesis enhances mTORC1 activity and promotes tumor progression. *Cell Metabolism* 29, 5 (2019), 1151–1165.e6. <https://doi.org/10.1016/j.cmet.2019.01.021>
- [4] Jing Feng, Jun-Ping Wang, Jian-Ran Hu, Ping Li, Pin Lv, Hu-Cheng He, Xiao-Wei Cheng, Zheng Cao, Jia-Jing Han, Qiang Wang, Qian Su, and Li-Xin Liu. 2025. Multi-omics Reveals the Associations Among the Fecal Metabolome, Intestinal Bacteria, and Serum Indicators in Patients with Hepatocellular Carcinoma. *World Journal of Gastroenterology* 31, 15 (Apr 2025), 104996. <https://doi.org/10.3748/wjg.v31.i15.104996>
- [5] David A. Guertin and Kathryn E. Wellen. 2023. Acetyl-CoA Metabolism in Cancer. *Nature Reviews Cancer* 23 (Mar 2023), 156–172. <https://doi.org/10.1038/s41568-022-00543-5>
- [6] Mohammad Amin Hemmati, Marzieh Monemi, Shima Asli, Sina Mohammadi, Behina Foroozanmehr, Dariush Haghmorad, Valentyn Oksenysh, and Majid Eslami. 2024. Using New Technologies to Analyze Gut Microbiota and Predict Cancer Risk. *Cells* 13, 1987 (Dec 2024). <https://doi.org/10.3390/cells13231987>
- [7] Karolina Kaźmierczak-Siedlecka, Ewa Stachowska, Marcin Folwarski, Katarzyna Przewłocka, Wojciech Makarewicz, and Ewa Bryl. 2021. The potential of gut microbiome as a non-invasive predictive biomarker for early detection of pancreatic cancer and hepatocellular carcinoma. *European Review for Medical and Pharmacological Sciences* 25, 23, 7275–7284. https://doi.org/10.26355/eurrev_202112_27421

- [8] Wenyu Luo, Shiqi Guo, Yang Zhou, Jingwen Zhao, Mengyao Wang, Lixuan Sang, Bing Chang, and Bingyuan Wang. 2022. Hepatocellular Carcinoma: How the Gut Microbiota Contributes to Pathogenesis, Diagnosis, and Therapy, Xiaofei Yu (Ed.). *Frontiers in Microbiology* 13. <https://doi.org/10.3389/fmicb.2022.873160>
- [9] S. Magnúsdóttir, D. Ravcheev, V. de Crécy-Lagard, and I. Thiele. 2015. Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes. *Cell Reports* 13, 12 (2015), 2651–2663.
- [10] T. Oliver and collaborators. 2023. TaxaHFE: Enhancing microbial classification through hierarchical feature engineering. *Bioinformatics* 39, 5 (2023), 1234–1245.
- [11] E. Oudah et al. 2018. Hierarchical feature engineering improves accuracy in microbiome-based machine learning models. *BMC Bioinformatics* 19, 256 (2018), 1–10.
- [12] Felipe P. Oñate, Juan D. Ramírez, and Jorge U. Marquardt. 2023. The Role of Adlercreutzia in Gut Microbiota and Human Health. *International Journal of Molecular Sciences* 24, 12 (2023), 12232. <https://doi.org/10.3390/ijms24122232>
- [13] Yotsawat Pomyen, Jittiporn Chaisaingmongkol, Siritida Rabibhadana, Benjarath Pupacdi, Donlaporn Sripan, Chidchanok Chornkrathok, Anuradha Budhu, Vajarabhongsa Budhisawasdi, Nirush Lertprasertsuke, Anon Chotirosniramit, Chawalit Pairojkul, Chirayu U. Auewarakul, Teerapat Ungtrakul, Thaniya Sricharunrat, Kannikar Phornphutkul, Suleeporn Sangrajang, Christopher A. Loffredo, Curtis C. Harris, Chulabhorn Mahidol, Xin Wei Wang, Mathuros Ruchirawat, and TIGER-LC Consortium. 2023. Gut dysbiosis in Thai intrahepatic cholangiocarcinoma and hepatocellular carcinoma. *Scientific Reports* 13. <https://doi.org/10.1038/s41598-023-38307-2>
- [14] Zhigang Ren, Ang Li, Jianwen Jiang, et al. 2019. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early detection of pancreatic cancer and hepatocellular carcinoma, Vol. 68. *BMJ*, 1014–1023. <https://doi.org/10.1136/gutjnl-2017-315084>
- [15] Harriet Rumgay, Jacques Ferlay, Catherine de Martel, Damien Georges, et al. 2022. Global, regional and national burden of primary liver cancer by subtype. *European Journal of Cancer* 161 (Mar 2022), 108–118. <https://doi.org/10.1016/j.ejca.2021.11.023>
- [16] Jakub Smajdor, Kamila Jedlińska, Rafał Porada, et al. 2023. The Impact of Gut Bacteria Producing Long Chain Homologs of Vitamin K₂ on Colorectal Carcinogenesis. *Cancer Cell International* 23 (2023), 268. <https://doi.org/10.1186/s12935-023-03114-2>
- [17] Fan Xv, Jiepeng Chen, Lili Duan, and Shuzhuang Li. 2018. Research Progress on the Anticancer Effects of Vitamin K₂. *Oncology Letters* 15, 6 (2018), 8926–8934. <https://doi.org/10.3892/ol.2018.8502>
- [18] Hao Zhang, Iwata Ozaki, Hiroshi Hamajima, Shinji Iwane, Hirokazu Takahashi, Yasunori Kawaguchi, Yuichiro Eguchi, Kyosuke Yamamoto, and Toshihiko Mizuta. 2011. Vitamin K₂ Augments 5-Fluorouracil-Induced Growth Inhibition of Human Hepatocellular Carcinoma Cells by Inhibiting NF- κ B Activation. *Oncology Reports* 25, 1 (2011), 159–166.
- [19] Ningning Zhang, Zeyu Wang, Jiayu Lv, et al. 2022. Characterization of Gut Microbiota and Exploration of Potential Predictive Model for Hepatocellular Carcinoma Microvascular Invasion. *Frontiers in Medicine* 9. <https://doi.org/10.3389/fmed.2022.836369>