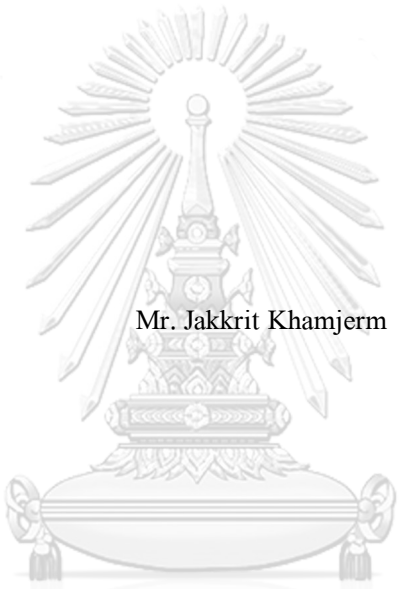
INTEGRATED MULTI-OMICS ANALYSIS OF GUT MICROBIOME AND HOST TRANSCRIPTOME TO IDENTIFY NOVEL BIOMARKERS FOR HEPATOCELLULAR CARCINOMA

Mr. Jakkrit Khamjerm



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Biomedical Engineering
Faculty Of Engineering
Chulalongkorn University
Academic Year 2023

การศึกษาความสัมพันธ์ระหว่างจุลินทรีย์ในลำไส้และการแสดงออกของยีนโดยการวิเคราะห์แบบมัลติโอมิกส์เพื่อหาตัวบ่งชี้ทางชีวภาพใหม่สำหรับโรคมะเร็งตับ

นายจักรกฤษณ์ ขำเจิม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมชีวเวช
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2566

Thesis Title	INTEGRATED MULTI-OMICS ANALYSIS OF GUT MICROBIOME AND HOST TRANSCRIPTOME TO IDENTIFY NOVEL BIOMARKERS FOR HEPATOCELLULAR CARCINOMA
By	Mr. Jakkrit Khamjerm
Field of Study	Biomedical Engineering
Thesis Advisor	Associate Professor THANARAT CHALIDABHONGSE, Ph.D.
Thesis Co Advisor	NATTHAYA CHUAYPEN, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF ENGINEERING

(Professor SUPOT TEACHAVORASINSKUN, Ph.D.)

THESIS COMMITTEE

..... Chairman

(Associate Professor DUANGDAO WICHADAKUL, Ph.D.)

..... Thesis Advisor

(Associate Professor THANARAT CHALIDABHONGSE, Ph.D.)

..... Thesis Co-Advisor

(NATTHAYA CHUAYPEN, Ph.D.)

..... Examiner

(Professor Pisit Tangkijvanich, M.D.)

..... External Examiner

(Associate Professor Teerapong Leelanupab)

จักรกฤษณ์ ขำเจิม : การศึกษาความสัมพันธ์ระหว่างจุลินทรีย์ในลำไส้และการแสดงออกของยีนโดยการวิเคราะห์แบบมัลติโอมิกส์เพื่อหาตัวบ่งชี้ทางชีวภาพใหม่สำหรับโรคมะเร็งตับ. (INTEGRATED MULTI-OMICS ANALYSIS OF GUT MICROBIOME AND HOST TRANSCRIPTOME TO IDENTIFY NOVEL BIOMARKERS FOR HEPATOCELLULAR CARCINOMA) อ.ที่ปรึกษาหลัก : รศ. ดร.ธนารัตน์ ชลิดาพงศ์, อ.ที่ปรึกษาร่วม : ดร.ณัฐธยาน์ ช่วยเพ็ญ

ความไม่สมดุลของจุลินทรีย์ในลำไส้มีความสัมพันธ์อย่างมากกับโรคมะเร็งตับชนิดปฐมภูมิหรือมะเร็งเซลล์ตับ (hepatocellular carcinoma; HCC) โดยตับกับลำไส้นั้นมีความเชื่อมโยงผ่านทางระบบไหลเวียนตับและลำไส้โดยผ่านทางแกนลำไส้และตับ (Gut-Liver axis) อย่างไรก็ตาม ความเข้าใจเกี่ยวกับความเชื่อมโยงกันของจุลินทรีย์ในลำไส้และการแสดงออกของโฮสต์ยีนยังคงมีจำกัด โดยวัตถุประสงค์ในการศึกษานี้มุ่งศึกษาไปยังความสัมพันธ์ระหว่างโปรไฟล์ของจุลินทรีย์ในลำไส้และโปรไฟล์การแสดงออกของโฮสต์ยีนในผู้ป่วยมะเร็งตับ ในการศึกษานี้ผู้วิจัยคัดเลือกผู้ป่วยมะเร็งตับ ที่มีสาเหตุมาจากการติดเชื้อไวรัสบีหรือซีจำนวน 17 ราย กลุ่มมะเร็งตับที่ไม่ได้มีสาเหตุมาจากการติดเชื้อไวรัสจำนวน 13 ราย และกลุ่มอาสาสมัครสุขภาพดีจำนวน 10 ราย และทำการตรวจสอบโปรไฟล์จุลินทรีย์ในลำไส้จากตัวอย่างอุจจาระ โดยใช้การวิเคราะห์การจัดลำดับของนิวคลิโอไทด์ของยีน 16S ribosomal RNA (16S rRNA) ด้วยเทคนิค next generation sequencing (NGS) และโปรไฟล์การแสดงออกของโฮสต์ยีนจากเซลล์เม็ดเลือดขาวชนิดโมโนนิวเคลียร์ โดยใช้การวิเคราะห์การจัดลำดับของอาร์เอ็นเอ ด้วยเทคนิค NGS เช่นเดียวกัน ชุดข้อมูลในแต่ละชุดได้รับการตรวจสอบและวิเคราะห์ร่วมกันเพื่อหาความสัมพันธ์ระหว่างชุดข้อมูลสองชุดโดยใช้เครื่องมือชีวสารสนเทศ นอกจากนี้ยังใช้โมเดลการเรียนรู้ของเครื่อง เพื่อระบุจุลินทรีย์ในลำไส้และยีนที่สามารถใช้สำหรับการวินิจฉัยมะเร็งตับ จากผลการวิเคราะห์สหสัมพันธ์ของเพียร์สันโดยใช้ข้อมูลจุลินทรีย์ในลำไส้จำนวน 268 แท็กซ่า และ ยีนจำนวน 6,137 ยีน พบว่าจุลินทรีย์ในลำไส้ 4 แท็กซ่า มีความสัมพันธ์กับการแสดงออกของโฮสต์ยีน 18 ยีน ซึ่งเป็น แบคทีเรียที่สังเคราะห์ไลโปโพลีแซ็กคาไรด์ และมีความสัมพันธ์กับยีนที่มีความเกี่ยวข้องกับการตอบสนองของระบบภูมิคุ้มกันของร่างกาย ซึ่งมีบทบาทในการส่งเสริมการพัฒนาของโรคมะเร็งตับ จากนั้นได้นำโมเดลการเรียนรู้ของเครื่องมาใช้ทดสอบประสิทธิภาพของการวินิจฉัยโรค พบว่า จุลินทรีย์ในลำไส้จำนวน 4 แท็กซ่า ได้แก่ *Eubacterium*, , *Eubacterium nodatum group*, *Lachnospiraceae AC2044 group* และ *Ruminococcus gnavus group* สามารถนำมาใช้เป็นตัวบ่งชี้ทางชีวภาพในการวินิจฉัยแยกผู้ป่วยมะเร็งตับที่ไม่ได้มีสาเหตุมาจากการติดเชื้อไวรัสตับอักเสบ กับผู้ป่วยมะเร็งตับที่มีสาเหตุมาจากการติดเชื้อไวรัสตับอักเสบ (AUC = 0.85, Sensitivity = 88%, Specificity = 80% and Accuracy = 86%) อย่างไรก็ตามโฮสต์ยีนที่พบดังกล่าวนั้น มีประสิทธิภาพในการแยกโรคได้ไม่ดีเท่าที่ควร จากการศึกษาครั้งนี้ชี้ให้เห็นว่าการเปลี่ยนแปลงของแบคทีเรียที่จำเพาะมีความสัมพันธ์กับการแสดงออกของโฮสต์ยีน ดังนั้นการปรับเปลี่ยนจุลินทรีย์และเพิ่มความสมดุลของจุลินทรีย์ในลำไส้ อาจมีส่วนช่วยในการชะลอการดำเนินโรค โดยเฉพาะอย่างยิ่งในผู้ป่วยมะเร็งตับชนิด HCC ที่ไม่ได้มีสาเหตุมาจากการติดเชื้อไวรัส

สาขาวิชา	วิศวกรรมชีวเวช	ลายมือชื่อนิสิต .....
ปีการศึกษา	2566	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6470111021 : MAJOR BIOMEDICAL ENGINEERING

KEYWORD: Hepatocellular carcinoma, Transcriptome, Gut microbiome, Diagnosis, Biomarkers

Jakkrit Khamjerm : INTEGRATED MULTI-OMICS ANALYSIS OF GUT MICROBIOME AND HOST TRANSCRIPTOME TO IDENTIFY NOVEL BIOMARKERS FOR HEPATOCELLULAR CARCINOMA. Advisor: Assoc. Prof. THANARAT CHALIDABHONGSE, Ph.D. Co-advisor: NATTHAYA CHUAYPEN, Ph.D.

An imbalance in gut microbiome is strongly linked to liver inflammation disease and hepatocellular carcinoma (HCC) via the gut-liver axis. However, the understanding of how gut microbiota interacts with the host gene expression is still limited. In this study, we aim to investigate the relationship between gut microbiome profile and transcriptomic profile in patients with HCC. In this study, 17 patients with viral-related HCC, 13 non-viral-related HCC, and 10 healthy controls were recruited. We investigated gut microbiome profile from fecal samples using 16S rRNA sequencing and host transcriptomic profile from the peripheral blood mononuclear cells (PBMCs) using RNA sequencing method. Individual datasets were examined and integrated for association analysis between two datasets using bioinformatic tools. Moreover, machine learning has been performed to detect HCC and then identify that bacterial and genes that can be used as diagnostics for HCC. Based on Pearson's correlation analysis, the interaction of 268 gut microbes and 6,137 genes were performed. We found that 4 genera of bacteria were associated with 18 host genes expression. In these interactions, these bacteria was related to lipopolysaccharide (LPS) production and the functional analysis of those genes was mainly involved in signal transduction and immune regulation. Finally, based on machine learning approach, 4 genera of bacteria including *Eubacterium*, *Eubacterium nodatum group*, *Lachnospiraceae AC2044 group* and *Ruminococcus gnavus group* were revealed to be diagnostic biomarkers in discriminating non-viral-related HCC from viral-related HCC (AUC = 0.85, Sensitivity = 88%, Specificity = 80% and Accuracy = 86%). However, the performance in differentiate the non-viral and viral-related HCC of host genes were not satisfactory. Our results suggested that alteration of the abundance of specific taxa was associated with specific host gene expression. The modulation of gut microbiota might improve gut homeostasis especially in patients with non-viral-related HCC.

Field of Study:	Biomedical Engineering	Student's Signature .....
Academic Year:	2023	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have supported me throughout the process of preparing this thesis proposal. Their guidance, encouragement, and contributions have been invaluable to me.

I am deeply thankful to my supervisor, Assoc. Prof. Thanarat Chalidabhongse, for their unwavering support and insightful guidance. Their expertise and constructive feedback have significantly shaped the direction of this research.

I wish to convey my gratitude for the collaborative dialogues and creative sessions shared with my co-advisor, Dr. Natthaya Chuaypen. These interactions have played a pivotal role in elucidating my concepts and elevating the holistic caliber of this proposal. Dr. Natthaya profound expertise, perceptive input, and unwavering commitment have significantly guided the trajectory of my research.

I would like to extend a special and heartfelt thank you to Prof. Pisit Tangkijvanich has gone above and beyond, dedicating their time, expertise, and insightful perspectives to guide me through the intricacies of this research endeavor. Their invaluable contributions have been a driving force behind the shaping of my ideas and the refinement of my approach.

I am also grateful to the members of my thesis committee, Assoc. Prof. Duangdao Wichadakul, Prof. Pisit Tangkijvanich, and Assoc. Prof. Teerapong Leelanupab, for their valuable suggestions and critical insights that have contributed to refining the research questions and methodology.

I extend my appreciation to the Center of Excellence in Hepatitis and Liver Cancer and member who have provided valuable discussions and a supportive environment to nurture my ideas.

My heartfelt thanks go to my family for their constant encouragement and understanding throughout this journey. Your belief in me has been a driving force in pushing me forward.

Lastly, I want to acknowledge the scholarship from the graduate School, Chulalongkorn University to commemorate the 72nd anniversary of his Majesty King Bhumibol Aduladej, the funding support from NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation, the countless researchers and authors whose work has laid the foundation for my study. Their contributions to the field have been instrumental in

shaping my understanding of the subject matter.

In conclusion, I am humbled by the support and contributions of all those mentioned above. This work would not have been possible without their assistance. However, any errors or omissions that might remain in this proposal are solely my responsibility. Thank you.

Jakkrit Khamjerm

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1 Introduction

1.1 Background of research

Liver cancer is currently one of the major health issues in the world with an anticipated incidence of more than 1 million cases by 2025 [1, 2]. Hepatocellular carcinoma (HCC) is the most common type of liver cancers and it has extremely high 90% mortality-to-incidence ratio. Hepatitis B and C infections are the leading factor of the development of HCC although alcoholic, non-alcoholic steatohepatitis (NASH), non-alcoholic fatty liver diseases (NAFLD), aflatoxin exposure and fluke infection can lead to the progression of HCC [3, 4].

Currently, the modalities recommended for surveillance of HCC are liver ultrasound combined with or without biological biomarkers, such as alpha-fetoprotein (AFP) every 6 months. AFP is the most commonly used biomarker in surveillance and diagnosis of HCC [5, 6]. It is considered positive if its level is higher than 20 ng/mL. However, its specificity and sensitivity are limited in the early-stage of HCC [7, 8].

According to 2012 from previous studies, the human body is composed of both visible and invisible components. The complete human body consists of the visible organs and invisible microorganisms, such as bacteria, fungi and viruses. In fact, invisible cells contain many more genes than visible cells [9]. The study of the human microbiome project began around 2,000. The purpose of this research was to characterize the microbial population, which inhabited the human body, and demonstrate how the microbial populations in the different parts of the body differed from each other [10]. Currently, the gut-liver axis has been the focus of this research. It refers to the bidirectional communication between liver and intestinal. Primary bile acids and antimicrobial molecules should be secret to the biliary tract by the liver. Molecules in the intestinal lumen that are supported by gut microbiomes can convert primary bile acid to secondary bile acid, which is not harmful. This procedure retains the balance of the liver and intestinal system, called eubiosis [11]. A stronger understanding of the factors affecting the manipulation of gut microbiome diversity on liver disease has been established from studies of the link between gut microbiota and metabolites [12].

A new paradigm in biomedical research to identify the genetic cause of human disorders has been accomplished by next generation sequencing (NGS) [13]. In previous studies, NGS was applied to identify novel genetic alterations and cancer genomes that drove tumor progression

[14]. The study in transcriptome profile can provide overall of genes expression and help understand the disease mechanism. Currently, most transcriptomic profile studies with RNA-sequencing technique use samples from tissues. However, few studies using liquid biopsy sample including peripheral blood mononuclear cells (PBMCs) are reported [15].

The use of integrated multi-omics of gut microbiome and host transcriptome in PBMCs of viral- and non-viral related-HCC patients as potential biomarkers has not been reported yet. We obtained paired fecal and blood specimen from the patients and healthy individuals. The objective of this study is to examine the association between gut microbiome and host gene expression data. Then, machine learning models were performed to analyze significant candidate bacteria or genes for diagnosis of HCC. Finally, we hope that this study will be able to find diagnostic biomarker for differentiating etiology of HCC and improve efficacy of diagnosis in HCC patients, especially in non-viral related-HCC patients.

1.2 Research questions

• Do the gut microbiota and host transcriptome profiles of HCC patients differ from healthy controls?

• Do the gut microbiota associate with host transcriptome profile in patients with HCC?

• Does the interaction of gut microbiota and host transcriptome represent as biomarkers for diagnosis of HCC

1.3 Objectives of work

• To investigate gut microbiome profiles of HCC patients and control groups.

• To investigate host gene profiles of HCC patients and control groups.

• To explore the interaction of the host-microbe in HCC using bioinformatic tools.

• To identify the biomarkers of host-microbe in diagnosis of patients with HCC.

1.4 Hypothesis

• Gut microbiome profiles of HCC patients may differ in comparison with those of healthy controls.

• Host transcriptome profiles of HCC patients may differ in comparison with those of healthy controls.

• Association of gut microbiota and host transcriptome might be used as diagnostic biomarkers for HCC.

1.5 Expected benefits

• The different between gut microbiome and transcriptomic profiles of Asian HCC patients who different lifestyle including habitat, heredities and dietary backgrounds.

• The association of gut microbiome and host transcriptome in HCC patients with viral and non-viral related, including the impact of the gut-liver axis on gut microbiome diversity and differential expression genes to clarify how gut microbiota influence the transcriptome profiles of HCC.

• The use of the novel non-invasive biomarkers from outcome of this study, together with the conventional biomarkers to benefit efficacy of diagnosis in HCC patients.

Chapter 2 Literature review

2.1 Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is the most prevalent type of primary liver cancers and has a high mortality rate from cancer. The mortality to incidence ratio is 0.91. It affects men 2.3 times more commonly than women while 72% of newly diagnosed cases are found in Asia [1, 2]. The most significant risk factors for the development of HCC continue to be chronic liver disease and cirrhosis whereas viral hepatitis and excessive alcohol consumption are ranked as the most relevant factors [3, 4].

Cirrhosis and/or HCC can develop from chronic viral hepatitis. The most two types of hepatitis that cause chronic hepatitis are hepatitis B and C. The double-stranded, circular DNA molecule known as the hepatitis B virus (HBV) has eight genotypes (A to H). In comparison to genotypes B and C, genotypes A and D are more prevalent in Asia and the Middle East [5]. Sexual contact, intravenous injections, and tainted blood transfusions are ways that hepatitis B is transmitted. The main source of HBV infection worldwide is vertical transfer from mother to fetus. Hepatitis B is a disease that affects 5% of people worldwide [16]. There are six distinct HCV genotypes that are isolated. Genotypes I, II, and III are more prevalent in Western countries, while type IV is more prevalent in the Middle East. 80% of HCV-positive patients proceed to chronic hepatitis, and 20% of HCV-positive patients eventually develop cirrhosis [6]. Similar to the emerging roles of NAFLD and metabolic syndrome in the development of HCC, alcoholism and consumption of foods contaminated with aflatoxin B1 are additional risk factors. The continued use of alcohol increases the risk of developing HCC. The level of alcohol taken over the course of a lifetime is correlated with the risk of liver disease while heavy drinking is being more associated with HCC than social drinking [17].

The HBV genotype can be classified into 8 genotypes (A to H) and four of these genotypes have been lately described with subgenotypes (A, B, C and F). The genotypes exhibit a clear geographic separation. The common prevalence of HBV genotype worldwide is genotype A and found in northwestern Europe, North America, and Africa [18]. Meanwhile, HBV genotype B and C are frequently found is Asia and Oceania [18]. Genotype D is widespread throughout the world but it is most prevalent in the Mediterranean region [18]. Genotype E is found in Africans on the West Coast of Africa and Madagascar on the east [19]. Genotypes F and H are only found

in the Amerindian communities of Central America [20] while genotype H is also discovered in California and Mexico [21]. Genotype G has only been isolated so far from HBV carriers in France, Germany, the United Kingdom, Italy, and the United States of America (USA) [22]. Most HBV genotypes found in Thailand are genotype C and genotype B accounting for 87.5% and 10.5%, respectively [23].

HCV genotypes can be classified into 6 genotypes (I to VI) and sub genotypes are approximately 25% of nucleotide sequence dissimilarity [24]. Genotypes I and III are the majority of infections worldwide and found mostly in East Asia. Genotypes II and IV are found in East Asia while Genotype IV is mostly found in North and Middle East Africa. Genotype V is only found in Southern and Eastern Africa [25]. The majority of HCV found in Thailand are genotype I, III and VI accounting for 28%, 31% and 41%, respectively [26].

2.2 Hepatitis B-related hepatocellular carcinoma

HCC is most frequently caused by chronic HBV infection worldwide. HBV is responsible for more than 50% of HCC cases globally and 70–80% of HCC cases in regions with a high HBV epidemiology. The mechanism of chronic hepatitis as a significant risk factor for HCC has been recognized in hepatic cells, activated immune response, cytokine release, inflammation and fibrosis. HBV-related HCC develops after 25 to 30 years of infection as a result of recurrent cellular regeneration and chronic inflammation. Another important risk factor for the development of HCC is cirrhosis carried upon with chronic hepatic injury. Cirrhosis is reported in 80–90% of HCC cases associated with HBV [27].

2.2.1 Hepatitis B infection

HBV infection is mainly transmitted by blood and semen. Three main transmission methods are mostly found. In regions with a high endemicity, most neonatal HBV transmission occurs between infected mothers and their newborn babies. In regions with a low endemicity, the risk of infection is high in the number of sex partners and men who have sex with men (MSM). The last major source of infection is inappropriate injections, blood transfusions, or hemodialysis [28].

2.2.2 Hepatitis B genome and structure

HBV is a member of the Hepadnavirus family. The infectious HBV virion has a diameter of 42 nm and contains 3.2 kb of partly double stranded rcDNA in a nucleocapsid (core) that is encapsulated in a lipid bilayer dotted with complexes of viral glycoproteins (Figure 1). A single copy of the viral genomic DNA and DNA polymerase are packaged in the nucleocapsid, which is made up of viral capsid proteins. There are three viral glycoproteins called "large" (L, LHBs), "middle" (M, MHBs), and "small" (S, SHBs) surface antigens that are present on the envelope membrane [29, 30].

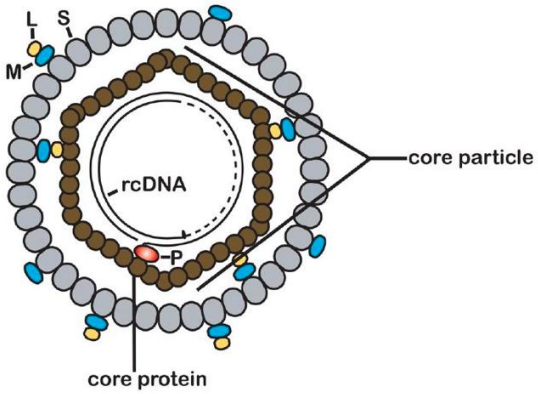


Figure 1 The mature HBV virion [30]

2.3 Hepatitis C-related hepatocellular carcinoma

Hepatitis C virus has demonstrated to be a major health concern due to the cause of liver cirrhosis and risk of developing liver cancer [31]. According to the current reports, the prevalence rate increased during the past decade to 2.8%, or more than 185 million infections worldwide [32].

2.3.1 Hepatitis C infection

HCV is mainly spread via percutaneous blood exposure associated with health procedures or through sharing contaminated injection equipment. Sexual and mother-to-infant transmission are also possible but less frequent. At present, receiving a tattoo in an unregulated setting, patient-to-patient transmission, and needle-stick injuries among healthcare workers remain risk factors for HCV transmission [33].

2.3.2 Hepatitis C genome and structure

HCV is a member of the Flaviviridae family, genus Hepacivirus [34]. HCV genome is a single stranded RNA of positive polarity containing 9.6 kb. HCV virions have an approximate 45 - 65 nm diameter and two anchoring envelope glycoproteins (E1 and E2) that are encased in a lipid bilayer [35]. The HCV genome contains 2 sections. The first is the untranslated (UTR) region and the second is the coding region (Figure 2) [36, 37].

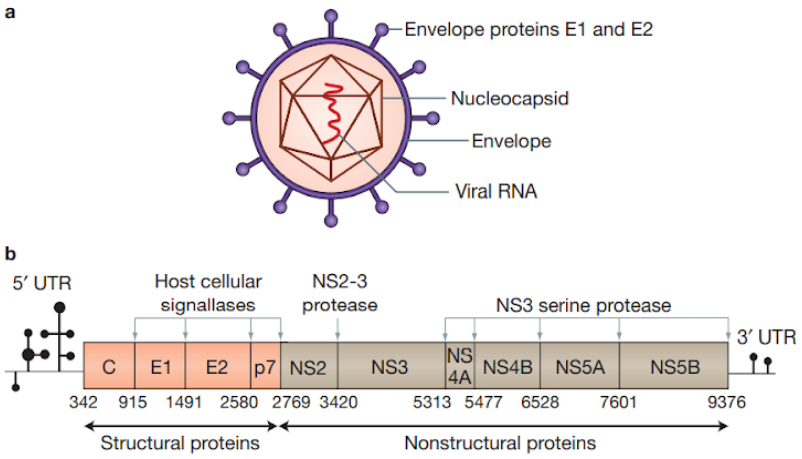


Figure 2 HCV virion and genome organization [37]

2.4 non-B non-C (NBNC)-related hepatocellular carcinoma

There are numerous HCC patients (5–20%) who test negative for the hepatitis C and hepatitis B virus infection indicators called non-B and non-C (NBNC) [38, 39]. The underlying liver diseases that contribute to NBNC-HCC vary including NAFLD, NASH, alcoholic liver disease, autoimmune hepatitis (AIH), primary biliary cirrhosis (PBC), primary sclerosing cholangitis (PSC) and aflatoxins. However, NAFLD (non-alcoholic fatty liver disease) is the most common cause of liver diseases. Patients with increased body mass index (BMI) and diabetes mellitus (DM) are associated with developing NAFLD. Moreover, NAFLD can lead to liver cirrhosis and HCC according to increased clinical evidence [40, 41]. In addition, metabolic syndrome increases the risk of HCC. The incidences of HCC associated with inflammatory and angiogenic alterations driven along with insulin resistance and fatty liver disease have increased [42, 43].

2.5 Diagnosis of HCC

More than 40 years ago, Alpha-fetoprotein (AFP) was identified as a marker for HCC and described as a way to identify preclinical HCC [7]. Increased AFP that indicates may be tumor. AFP cutoff level of 10 to 20 ng/mL was reported to have a sensitivity and specificity of 80% and 60%, respectively [8]. Even with advanced disease, tumors that had normal AFP levels at the time of diagnosis frequently remained stable. So, the diagnosis challenge was based on only the AFP level. Liver biopsy is the gold standard for diagnosis of HCC. Unfortunately, it has many problems involved including invasive methods, causing pain, anxiety and discomfort to patients. Currently, more sensitivity techniques have been used, such as computerized tomography scan (CT) and magnetic resonance imaging (MRI) [44]. Even though imaging techniques have been recommended as the current guidelines for the diagnosis of HCC, their disadvantages include cost, radiation exposure, and the need for iodinated contrast [45].

Currently, liquid biopsy including nucleic acid, circulating tumor cells (CTCs) and extracellular vehicles (EVs) refer to molecular analysis and release into the bloodstream or other body fluids [46]. Therefore, this method has shown encouraging results for several cancer-related applications including non-invasive biomarkers for prognostic [47-49]. At present, peripheral blood mononuclear cells (PBMCs) can be demonstrated for alteration of total RNA representing cancer-induced genes that can serve as a new HCC prognostic and diagnostic marker [50-52].

2.6 The gut microbiota and hepatocellular carcinoma

Physiological relationship between intestinal tract and liver has been called the “gut-liver axis” (Figure 3) [12, 53]. An effect of metabolite in the intestinal on the liver is thought to play a key role in the development and progression of HCC [54]. Currently, gut microbiome can be suggested as a non-invasive biomarker for diagnosis of HCC [55]. Classification of disease severity by gut microbiome can be used for targeted and personalized treatment as well as being used as an indicator of the response to cancer [11, 56].

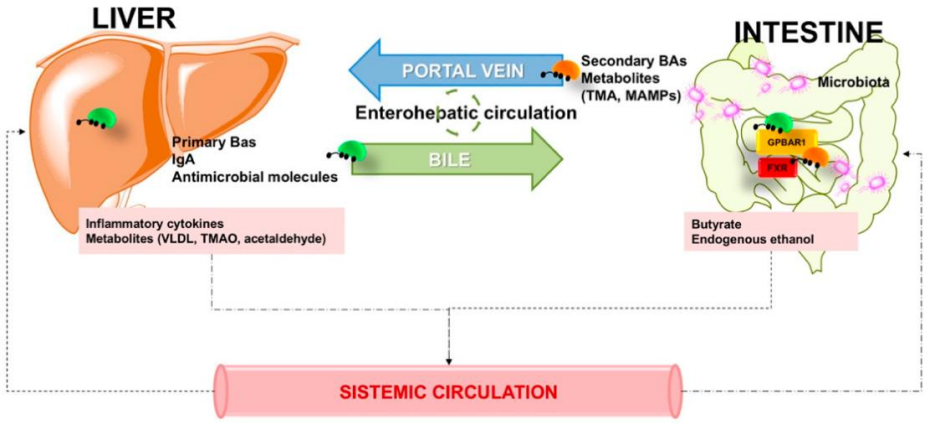


Figure 3 The communication between the liver and the gut is bidirectional [11]

Major gut microbiome metabolite product during the fermentation of polysaccharide is short chain fatty acids (SCFAs). For example Butyrate, which is mainly produced by Firmicutes phylum. Butyrate is essential role in an immunity activity and improved function intestinal barrier [57]. In a recent study of characterizing gut microbiome in hepatocellular carcinoma (HCC) patients with different stages and evaluating potential of microbiome to non-invasive biomarker for HCC, the results showed microbial diversity increased in liver cirrhosis and early HCC. Actinobacteria increased more in early HCC than liver cirrhosis, and Butyrate production decreased. Furthermore, to identify microbial biomarkers and construct HCC classifiers by the Random Forest model, the results showed an area under the curve of 80.64% between early HCC and non-HCC. This study has the strong diagnostic potential for early HCC and advanced HCC because it was validated in the HCC group from Northwest and Central China [58]. In another study of exploring what features of gut microbiota are associated with cirrhosis hepatocellular carcinoma (HCC) and non-alcoholic fatty liver disease (NAFLD), the whole population had three groups (cirrhosis with HCC-group, cirrhosis with non-HCC-group and healthy control-group). The results showed high abundance levels of Enterobacteriaceae and Streptococcus with low abundance levels of *Akkermansia* in the cirrhosis group. Meanwhile, it showed high abundance levels of *Bacteroides* and Ruminococcaceae with low abundance levels of *Bifidobacterium* in the HCC group. Moreover, the study explored intestinal permeability, inflammatory status and circulating mononuclear cells by cell assay. They constructed a model correlation of these features of HCC progression, the results founded correlation between *Akkermansia*, *Bacteroides* and *Bifidobacterium* with calprotectin. This study suggests gut microbiota from patients with

cirrhosis and NAFLD are significantly correlated with systemic inflammation in the process of hepatocarcinogenesis. It is unclear from this study's research gap if these alterations can vary depending on the disease's stage or if they may be linked to certain tissue or metabolic changes [59]. However, there is limited research on the association of gut microbiome and viral -HCC or non-viral-HCC.

2.7 16S rRNA sequencing for gut microbiome

The most widely used genetic marker has been the 16S rRNA gene sequence, which has been used in bacterial taxonomy and phylogeny research. The length of the 16S rRNA gene is enough for informatics approx. 1,500 base pairs to represent in almost all bacteria. Bacterial genome database being used with 16S rRNA sequencing to identify bacteria composition. Another difficult challenge is choosing primers that would specifically target specific 16S rRNA gene regions for bacterial taxonomy characterization. Several different 16S rRNA gene variable regions have been targeted in studies of gut microbiome, including V3, V4 and V3-V4. In Chen Z et al. primer pairs targeting the 16S rRNA gene V1-V2, V3-V4, and V4 regions was performed to profile the community of gut microbes. They discovered a higher alpha diversity and richness [60].

2.8 Transcriptomic profile in hepatocellular carcinoma

Genome-wide mapping of gene expression in tissue has been used for identifying biomarkers for diagnosis, prognosis, and new treatments in various diseases, especially cancer [13]. However, gene expression data based on microarray technique did not provide sufficient insight. RNA-sequencing is currently capable of evaluating changes at the molecular level that are related to disease pathogenesis [14].

At present, transcriptomic profile has the report with RNA -sequencing technique in tissues or liquid biopsy, for example peripheral blood mononuclear cells (PBMCs). In this research, the profiles of long non-coding RNAs (lncRNAs) obtained from PBMCs of HCC patients. The results showed gene expression levels of three up-regulating genes, MIR4435-2HG, SNHG9 and lnc-LCP2-1 and one down-regulating gene, lnc-POLD3-2. Functions of these genes are reported to have an association with carcinogenesis and immune response [15]. Moreover,

most previous studies in patients caused by hepatitis B or C virus infection and hepatocellular carcinoma at early stage could not be a measure for biomarker. In particular, only few studies have been conducted in patients with non-viral related hepatocellular carcinoma (NBNC-HCC). Consequently, research at the transcriptional level will support current data and provide a molecular perspective on the disease progression.

2.9 Machine learning in precision medicine

Machine learning-based big data analysis offers a number of benefits for integrating and analyzing a large amount of complex health-care data [61]. In the previous report, machine learning has been used to analyze biological data at various levels, including DNA, RNA and protein, as well as data from bacteria, such as gut microbiome. Integration of all data has been called “multi-omics" analysis [62]. The previous study was to find association between gut microbiome and host transcriptome in hepatitis B related with HCC patients. Moreover, this study used the models of Random Forest and Support Vector Machine model to further confirm gut microbiota's ability to predict clinical outcomes. The analysis of integration analysis between gut microbiota and host transcriptome showed 3 bacteria (*Bacteroides*, *Lachnospiraceae incertae sedis* and *Clostridium XIVa*) increased with non-small HCC and had relation with 31 genes with progression of cancer. Furthermore, the results showed the potential of gut microbiota for predicting clinical outcome yielding area under the curve at 81% [63]. However, the previous study was conducted in tissues samples. Therefore, the development of machine learning can be used as a tool for diagnosis and making treatment decisions more effective for patients. It is expected that the combination of machine learning with omics data from the same HCC patients can be used for determining patients who are more susceptible to develop liver cancer and allowing patients have a better quality of life in the future.

Chapter 3 Research Methodology

3.1 Research workflow

All samples were obtained from Chulalongkorn Memorial Hospital, Thailand and were used for gut microbiome and host gene expression analysis. Correlation based analyses were conducted to uncover microbe-associated genes, identify microbial markers, and develop HCC classifiers using machine learning model. The study integrated gut microbiome, transcriptome, and clinical data for comprehensive insights.

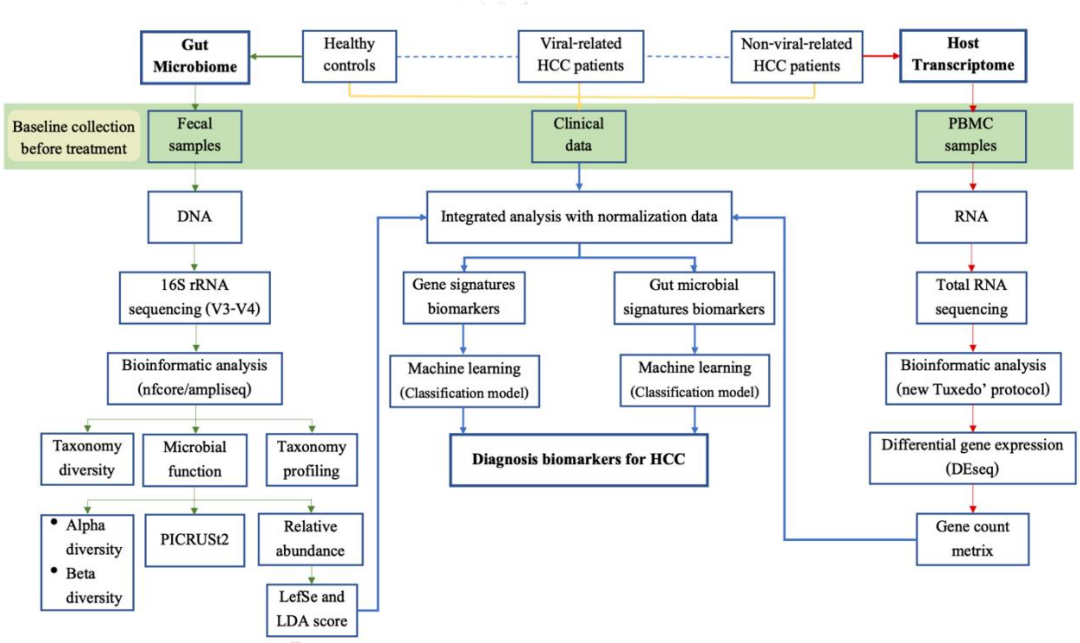


Figure 4 Research workflow. A total of fecal and blood samples from Chulalongkorn Memorial Hospital, Bangkok, Thailand were collected. DNA was extracted from fecal samples to characterize gut microbiome. RNA was extracted from blood samples to investigate host gene expression. Based on gut microbiome, transcriptome and clinical data, correlation-based analysis was performed to discover microbe-associated gene, identify microbial markers, and construct HCC classifier by machine learning model.

3.2 Experiment design

Fecal and blood samples from healthy control and HCC patients were obtained. DNA from fecal samples were characterized for the gut microbiome profile, while RNA from blood samples were analyzed for host gene expression profile. We aimed to integrated analysis of gut microbiome, transcriptome, and clinical data for the discovery of microbes-associated genes and microbial markers. Then, machine learning was performed in building the hepatocellular carcinoma (HCC) classifiers model.

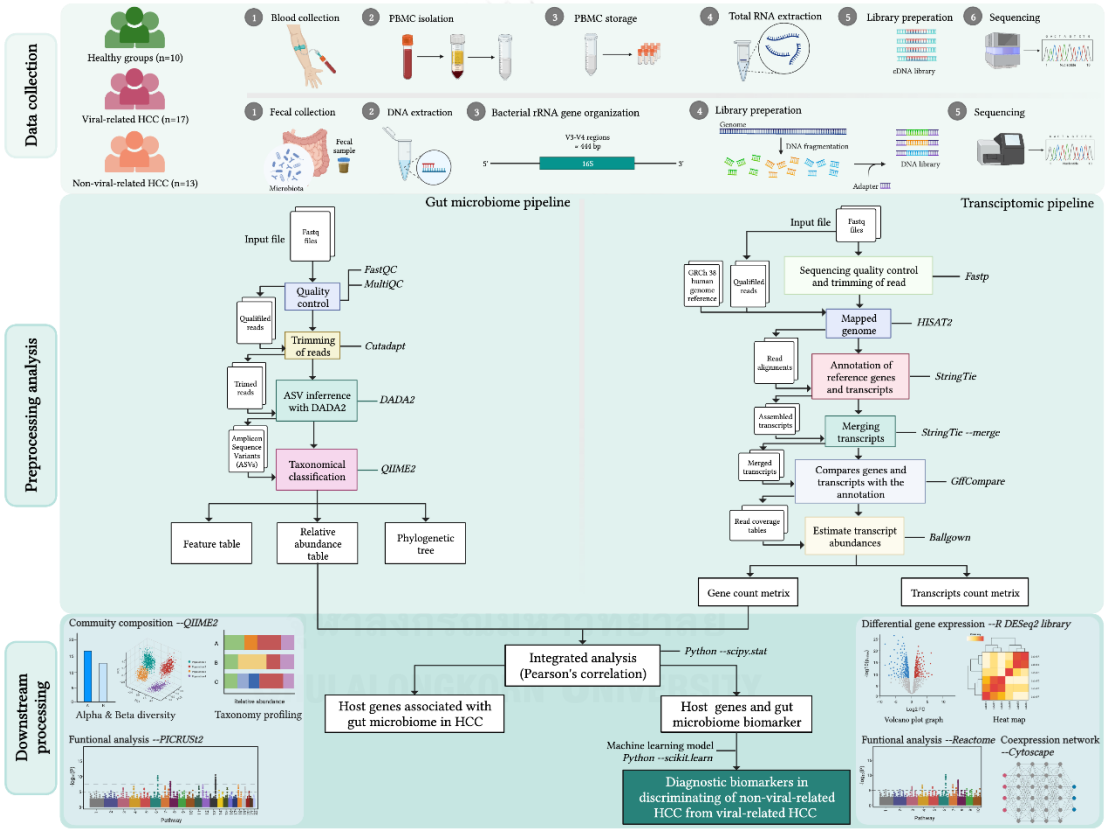


Figure 5 Experimental design in detail

3.3 Sample size calculation

This research was the case-control study, which are compared the gut microbiome and gene expression of patients with viral related-HCC and non-viral related-HCC. Sample size calculation was determined using Statulator provided at https://statulator.com/SampleSize/ss2P.html based on a reference study of Epidemiology and

Survival of Hepatocellular Carcinoma in the Central Region of Thailand from 2007 to 2012, the results was found that 5,929 patients were diagnosed with hematologic malignancy and 308 (5.19%) had final diagnosed with HCC [64]. The defined formula was shown in Figure 6.

$$n_{case} = \left[\frac{z_{1-\frac{\alpha}{2}} \sqrt{\bar{p}\bar{q}\left(1+\frac{1}{r}\right)} + z_{1-\beta} \sqrt{p_1 q_1 + \frac{p_2 q_2}{r}}}{\Delta} \right]^2$$

$$p_1 = P(exposure|case), q_1 = 1 - p_1$$

$$p_2 = P(exposure|control), q_2 = 1 - p_2$$

$$\bar{p} = \frac{p_1 + p_2 r}{1+r}, \bar{q} = 1 - \bar{p}, r = \frac{n_{control}}{n_{case}}$$

Figure 6 Formula for sample size calculation

P (exposure | case) = 0.0519 P (exposure | control) = 0.35

Ratio (control: case) = 1 α = 0.05 β = 0.20

n_{case} = 31 for each group

With the assuming of 5% of subjects in the reference population was the factor of interest. Therefore, the study would require a sample size of 32 for each group to achieve a power of 80% for detecting a difference in proportions of 0.20 between the two groups (test – reference group) at a two sided p-value of 0.05 [65].

3.4 Participant information

3.4.1 For healthy control group

The control group consisted of healthy volunteers who had no metabolic syndrome and liver diseases. The consent forms, which were completed by all participants before their samples were collected, was approved by the Institute Ethics Committee of the Chulalongkorn University Faculty of Medicine (IRB No.108/60 and IRB No.312/64).

- **Inclusion criteria for healthy control group**

Male and female Thai patients above or equal to 18 years old. Body mass index (BMI) and serological tests (including the detection of hepatitis B surface antigen and hepatitis C virus antibody) results were in the normal range. The healthy group had no evidence of liver disease and underlying history of metabolic syndromes.

- **Exclusion criteria for healthy control group**

The exclusion criteria for healthy control included hypertension, diabetes, obesity, metabolic syndrome, irritable bowel syndrome (IBD), non-alcoholic fatty liver disease and liver cirrhosis. Additionally, people, who had taken probiotics or antibiotics within the four weeks before enrollment, were excluded.

3.3.2 For HCC patient group

Total of HCC patients were diagnosed using the international guidelines at the King Chulalongkorn Memorial Hospital, Bangkok, Thailand. The consent forms, which were completed by all participants before their samples were collected, was approved by the Institute Ethics Committee of Faculty of Medicine, Chulalongkorn University (IRB No. 0371/66).

- **Inclusion criteria for HCC patient group**

Male and female Thai patients above or equal to 18 years old. The diagnosis of HCC patients was confirmed by computed Tomography (CT) and magnetic Resonance Imaging (MRI) regarding the clinical guideline of the American Association for the Study of Liver Diseases (AASLD). Patients with metabolic illnesses, such as hypertension, dyslipidemia, and type 2 diabetes were included. Additionally, patients, who previously had HCV or HBV infections and went on to develop HCC, were included.

- **Exclusion criteria for HCC patient group**

The exclusion criteria included patients with Intrahepatic cholangiocarcinoma, prior anticancer therapy, and participants missing clinical information or clinical outcome data. Moreover, people, who had taken probiotics or antibiotics within the four weeks before enrollment, were also excluded.

3.4 Sample collection

3.4.1 Fecal sample collection

Participants received guidance on how to employ the suitable fecal collection method based on the standard operating procedures (SOPs) [66]. A DNA/RNA Shield™ - Fecal Collection tube (Zymo Research Corp.) containing the reagent, which could preserve microbial nucleic acids and inactivate pathogens from fecal samples, was provided to the participants. In the laboratory, the samples were immediately stored at -80 °C until further experiment was required.

3.4.2 Blood sample collection

Blood specimens with an approximate size of 3 ml were collected in an EDTA tube from healthy control and HCC patients before performing chemoembolization treatments at King Chulalongkorn Memorial Hospital, Bangkok, Thailand between 2019 to 2021. Fresh EDTA blood specimens were used to isolate peripheral blood mononuclear cells (PBMCs). PBMCs were isolated at 2,500 rpm for 15 minutes at room temperature and then washed 2 times with PBS. The isolated PBMCs were suspended in PBS for 1 ml and stored at −80 °C until further experiment was required.

3.4.3 Clinical collection

Clinical characteristics data of all participants were collected before performing chemoembolization treatments at King Chulalongkorn Memorial Hospital, Bangkok, Thailand from hospital information system (HIS) including gender, age, body mass index (BMI), liver biochemistry, serological test, liver function, renal function, electrolyte, radiomic data, Child-Pugh classification, staging of HCC classified by the Barcelona Clinic Liver Cancer (BCLC), history underlying and overall survival times.

3.5 Fecal sample for DNA extraction

DNA was extracted from a 1ml frozen aliquot of each stool sample using ZymoBIOMICS™ DNA Miniprep kit (Zymo Research Corporation). DNA extraction using a bead beating system to complete homogenization/disruption of the microbial cell walls and accurate microbial DNA analysis. DNA concentration and purity were measured by DeNovix™ UV-Vis spectrophotometer and stored at -20 °C until further experiment was required. Moreover, we performed to amplify hypervariable region of bacterial genes (V3-V4 region) by polymerase

chain reaction (PCR) for confirmation. The PCR conditions started with the initial activation at 95°C for 2 minutes, the denaturation step at 95°C for 30 seconds, then the annealing step at 53°C for 40 seconds, the extension step at 72°C for 60 seconds and 40 cycles of amplification were recommended. The final step was the final extension at 72°C for 10 minutes. The PCR products can be evaluated by agarose gel electrophoresis before sending to 16S RNA sequencing distributor.

3.5.1 16S rRNA sequencing

The hypervariable V3-V4 region of the 16S rRNA gene (341F/785R) was targeted by a primer set used to amplify the extracted DNA samples (Table 1). Amplicon sequencing is a highly focused strategy that enables researchers to examine genetic diversity in certain genomic regions. In this procedure, Amplicon-based 16S rRNA was examined by ZymoBIOMIC®. Target sequencing of the DNA sample and amplification was performed using the Quick-16S™ NGS Library Prep Kit (ZymoResearch, CA) and real-time PCR technique, respectively. DNA clean and concentration by concentrator ™ (ZymoResearch, CA) were selected. DNA integrity was examined by TapeStation® (Agilent Technologies, USA) for library quantification. Positive control were used from ZymoBIOMIC® Microbial Community DNA standard. The final library will be sequenced on Illumina® MiSeq™ platform.

Table 1 Primer sequence for 16S rRNA sequencing

Primer name	Primer sequence	Amplicon size	Reference
V3-V4 region	5'-CCTACGGGNGGCWGCAG-3' 5'-CCTGCCTTTGCAATRTCIACRAANGC-3'	444 bases pair	[67]

3.5.2 Data preprocessing and analysis

Raw read data from Illumina® MiSeq™ platform of each sample following from nfcore/ampliseq analysis pipeline (doi: 10.5281/zenodo.1493841) (Figure 7) [68, 69]. The first step of ampliseq pipeline is to preprocess the data including FastQC and Cutadapt tools for sequencing quality control and trimming of read (primer and adapters). The output containing report quality metrics and summary of read numbers that pass Cutadapt tool. The next step to infer amplicon sequence variants (ASVs) using DADA2 tool. DADA2 reduces sequence errors

and dereplicates sequences by quality filtering, denoising and PCR chimera removal. In addition, DADA2 resolves variations of as little as one nucleotide and infers sample sequences exactly which is an advantage over traditional operational taxonomic units (OTU). The output containing fasta file with ASV sequences and counts for each ASV sequences. The next step for taxonomy classification was performed using the SILVA of 99% 16S rRNA gene reference database [70]. ASV sequences and counts data as produced before with DADA2 tool are imported into QIIME2 tool for taxonomic classification aligning with the reference database. The output contains tab-separated absolute abundance table at the taxa level. Moreover, QIIME2 tool can provide relative abundance tables using total sum scaling normalization (TSS) for various taxonomic levels as the final data for future downstream analysis and visualization. Barplot, diversity analysis (alpha and beta diversity) using various methods and performs pairwise comparison of groups of samples. PICRUSt2 is software for predicting the functional potential of a bacterial community based on marker gene sequences. Functional usually refers to several gene family databases are supported by default including the Kyoto Encyclopedia of Genes and Genomes (KEGG), orthologs (KO), Enzyme Classification (EC) numbers and MetaCyc ontology are among the features that PICRUSt2 could be capable of accomplishing (Figure 8) [71] The differentially abundant taxa was assessed using the Linear Discriminant Analysis Effect Size (LEfSe) method (https://huttenhower.sph.harvard.edu/galaxy/).

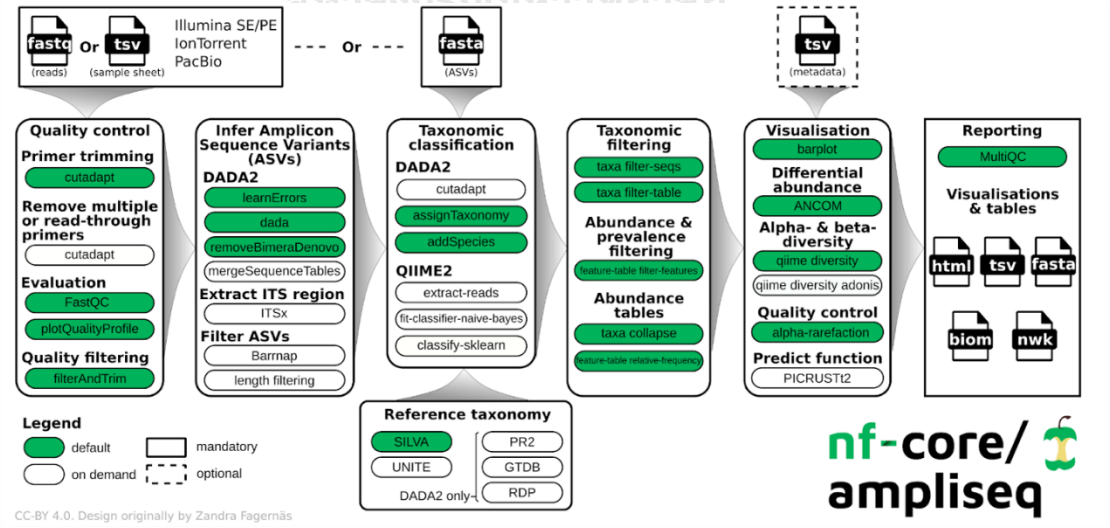


Figure 7 nf-core/ampliseq bioinformatics analysis pipeline [69]

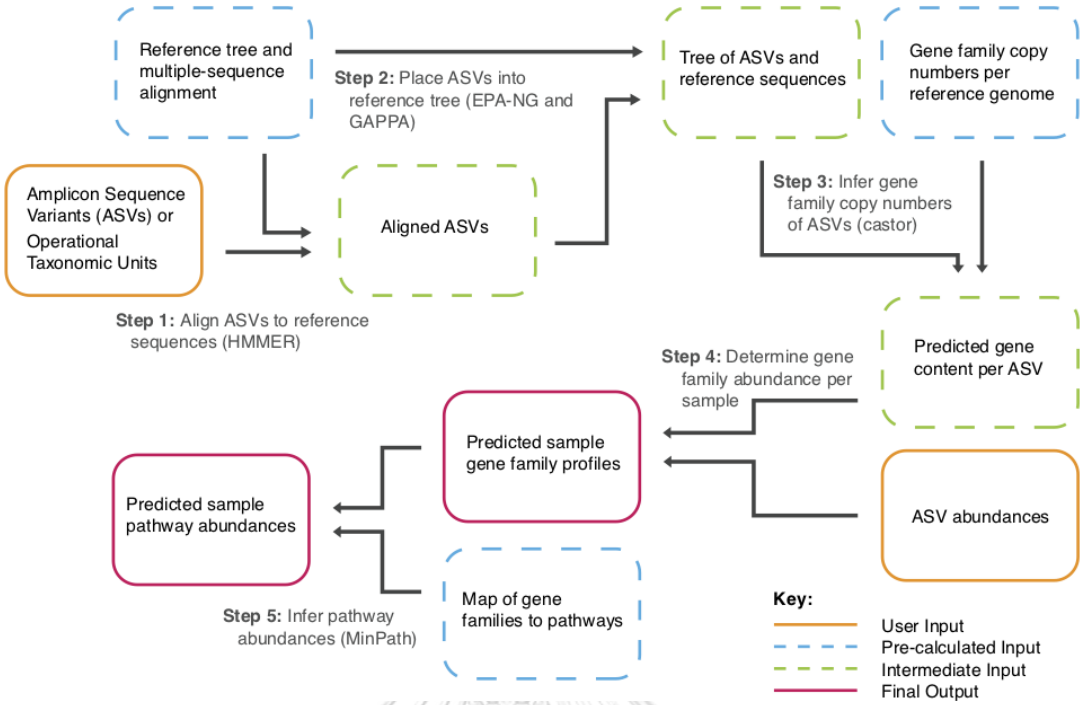


Figure 8 PICRUSt 2.0 Flowchart [71]

3.6 Blood sample for RNA extraction

Total RNA were extracted from PBMCs sample using TRIzol LS reagent (Invitrogen, USA) based on manufacturer's protocol. RNA concentration and RNA integrity were measured by RNA integrity by Qubit® 4 fluorometer (Invitrogen, USA) and TapeStation® (Agilent Technologies, USA), respectively.

3.6.1 Total RNA sequencing

Next generation sequencing library preparations were constructed according to the manufacturer's protocol (NEBNext® Ultra™ RNA Library Prep Kit for Illumina®). The poly(A) mRNA isolation was performed using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) or Ribo-Zero™ rRNA removal Kit (illumina). The mRNA fragmentation and priming was performed using NEBNext First Strand Synthesis Reaction Buffer and NEBNext Random Primers. First strand cDNA was synthesized using ProtoScript II Reverse Transcriptase and the second-strand cDNA was synthesized using Second Strand Synthesis Enzyme Mix. The purified double-stranded cDNA by AxyPrep Mag PCR Clean-up (Axygen, USA) was then treated with End Prep Enzyme Mix to repair both ends and add a dA-tailing in one reaction, followed by a T-A ligation to add adaptors to both ends to purified double-stranded cDNA. Size selection of

Adaptor-ligated DNA was then performed using AxyPrep Mag PCR Clean-up (Axygen, USA), and fragments of ~360 bp (with the approximate insert size of 300 bp) were recovered. Next, each sample was amplified by PCR for 11 cycles using P5 and P7 primers, with both primers carrying sequences which can anneal with flow cells to perform bridge PCR and P7 primer carrying a six-base index to allow multiplexing. The PCR products were later cleaned up using AxyPrep Mag PCR Clean-up (Axygen), validated using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and quantified by Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Then libraries with different indices were multiplexed and loaded on an Illumina HiSeq instrument according to manufacturer's instructions (Illumina, USA). Sequencing was carried out using a 2x150bp paired end (PE) configuration. Image analysis and base calling were conducted by the HiSeq Control Software (HCS) + OLB + GAPipeline-1.6 (Illumina) on the HiSeq instrument.

3.6.2 RNA-seq data preprocessing and analysis

RNA sequencing was analyzed based on 'new Tuxedo' protocol (Figure 9) [72]. Raw read data from Illumina Hiseq was performed using Fastp tool (version 0.21.1) for check quality, remove adapter, and remove for low quality sequence [73]. Sequencing reads were aligned using HISAT2 (version 2.1.0) with human reference sequence (Illumina GRCh38) [74]. StringTie tool (version 2.1.6) was used for alignment data to map for efficient transcript assembly and quantitation of RNA-Seq data [75]. Differential gene expression was analyzed with DESeq2 comparison between HCC subgroups. Total RNAs possessing a read count ^{3} 20 in ^{3}5 samples were chosen for subsequent analysis. A hierarchical cluster analysis of differentially expressed genes (DEGs) was performed to explore the expression pattern of genes in viral-related HCC and non-viral-related HCC groups. Specific gene expression that would be up-regulated and down-regulated were used at 1.5-fold change and P-value < 0.05.

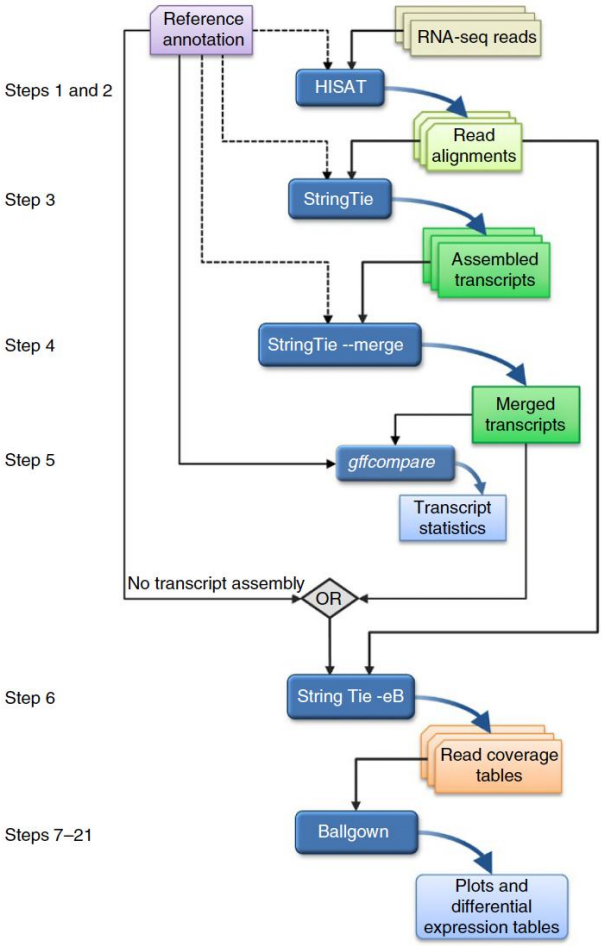


Figure 9 new Tuxedo protocol. RNA-seq read are mapped for each sample to the reference genome (Steps 1 and 2). The transcripts in each sample are assembled and quantified with StringTie (Step 3). After assembled, transcripts are merged together and creates a uniform set of transcripts for all samples (Step 4). The gffcompare program was used to compares the genes and transcripts with the annotation and reports statistics on this comparison (Step 5). The Ballgown tool provides functions to organize, visualize, and analyze the expression measurements for assembled transcripts (Step 6-7) [72].

3.7 Association between ASVs and differential gene expression

ASVs abundance and differential gene expression level were performed in correlation with Pearson's correlation coefficient for each pair ASV-gene across all samples. ASV that was presented in 10% of all samples was eliminated to decrease the computational load and minimize contingency. Gene expression values were calculated using DESeq2, such that each gene was

assigned a reliable fold change. Specific gene expression that would be up-regulated and down-regulated were used at 1.5-fold change and P value < 0.05. The statistical significance of each ASV-gene pair was determined by P-value < 0.05 and a false discovery rate (FDR) < 0.1. The GO enrichment analysis was performed based on Metascape [76]. The function analysis was performed based on Reactome pathway database (https://reactome.org) [77]. Reactome web base offers a complete range of functional annotation tools to help researchers comprehend the biological significance of lengthy gene lists.

3.8 Microbial and gene-based biomarker discovery for diagnosis

Model construction was performed on the relative abundance discriminating non-viral-related HCC from viral-related HCC using machine learning model. To address outliers in the features, we normalized numerical variables and encoded categorical features for classification purposes. To address the imbalance between the HCC subgroup datasets, we employed the Synthetic Minority Over-sampling Technique (SMOTE). This method involves generating synthetic instances of the minority class by interpolating between feature vectors of existing minority class examples. A total of 263 gut microbial taxa and 6,137 genes were considered for feature selection. Various techniques, such as Correlation-based Feature Selection were employed on all datasets corresponding to different HCC subgroups. Then the best classifier algorithm among Support Vector Machine, Random Forest and Logistic Regression was chosen based on their performance in classifying using features of gut microbiome and gene expression. This study applied a fundamental machine learning technique, the train-test split, which involved dividing the dataset into training and testing sets to evaluate the model's generalization to unseen data. The data was partitioned, with a test set comprising 30% of the total dataset. To assess the performance of each classifier model in K-fold cross validation, It allows the use of all the available data for both training and validation to produce a more robust estimate of the model's performance [78] and we employed metrics such as area under the curve (AUC), sensitivity, specificity, and accuracy. Moreover, we optimize hyperparameters using GridSearchCV, a method that systematically explores a predefined grid of hyperparameter values for a given model. Through cross-validation, it evaluates each combination to identify the optimal set of hyperparameters that achieves the best performance for the model.

3.9 Statistical analysis

Statistical analysis was performed using SPSS V28.0 (SPSS Inc, USA) and GraphPad Prism V8.0 (GraphPad software, USA). Chi's square or Fisher's exact test for categorical data and the Mann-Whitney U test for non-parametric values was used to compare the continuous data between the two groups. Microbiome diversity analysis was performed using pairwise comparisons of groups of samples. More than two groups were compared using ANOVA test. Statistical significance was defined with P-value < 0.05.

3.10 Ethical consideration

The Helsinki Declaration and Good Clinical Practice for the involvement of human subjects were followed in the study protocol's execution. Before fecal and blood samples were collected, each subject completed the informed consent forms, which were reviewed and approved by the Institute Ethics Committee of the Chulalongkorn University Faculty of Medicine (IRB No.108/60 and IRB No.312/64). The study was approved by the Institutional Review Board (IRB) of the faculty of Medicine, Chulalongkorn University, Bangkok, Thailand (IRB No.0371/66)

3.11 Expected benefit and application

In this study, data of gut microbiome and host transcriptome derived from healthy volunteers and HCC patients in Thailand population. The data describe daily life and nutrition in each group of samples. This study will comprehensively identify gut microbiomes and describe microbial diversity and correlation networks of gut microbiota. For RNA sequencing, the data provide differential gene expression, of which transcriptional profiles can be investigated. Understanding associations between two data sets provides new insights to explore the connections of gut microbiome and host transcriptome for human biomarker discovery. Additionally, machine learning models based on gut microbiome data can be used as diagnostic biomarkers in discriminating non-viral-related HCC from viral-related HCC.

Chapter 4 Result

4.1 Participant information

All of 30 patients with HCC and 10 healthy volunteers were enrolled in the study. From cause of HCC patients by hepatitis B virus, hepatitis C virus and non-B-non-C (NBNC) or non-viral-related HCC. Clinical characteristics of these group including healthy volunteers versus patients with HCC and viral-related HCC (n=17) versus non-viral-related HCC (n=13) were generally matched. BMI, Platelet, Albumin, AST, ALP, AFP, Maximum size and BCLC stage, suggesting that there was no significant confounding factors affecting group discrimination between comparing group (Table 2).

Table 2 Clinical characteristics summary of all participant

Clinical parameter	**Healthy (n=10)**	**Patients with HCC (n=30)**	***P-value***	**Patients with HCC (n=30)**		
				Viral-related HCC (n=17)	**Non-viral-related HCC (n=13)**	***P-value***
Age	34.3±10.3	65.3±10.1	<0.001*	61.1±9.2	70.8±8.6	0.007*
Gender			<0.001*			<0.001*
• Male	5(50%)	28(93.3%)		17(100%)	11(84.6%)	
• Female	5(50%)	2(6.7%)		0(0%)	2(15.4%)	
BMI	22.0±3.6	25.5±5.0	0.053	25.2±4.9	25.8±5.2	0.750
Platelet				128.8±59.1	174.0±89.4	0.108
Albumin				3.6±0.6	3.7±0.6	0.541
AST				66.1±57.1	46.5±28.9	0.267
ALT				51.3±31.1	28.2±11.6	0.017*
ALP				98.3±37.1	136.7±96.4	0.142
AFP				4476.6±17980.2	416.6±911.3	0.950
Total mass						0.045*
• 1				9(52.9%)	7(53.9%)	
• 2				4(23.5%)	1(7.7%)	
• >3				4(23.5%)	5(38.4%)	
Maximum size				4.8±4.2	6.2±6.3	0.450

Clinical parameter	Healthy (n=10)	Patients with HCC (n=30)	*P-value*	Patients with HCC (n=30)		*P-value*
				Viral-related HCC (n=17)	Non-viral-related HCC (n=13)	
Cirrhosis				15(88.2%)	12(92.3%)	
BCLC stage						0.122
• 0-A				8(47.1%)	6(46.2%)	
• B				7(41.2%)	4(30.8%)	
• C				2(11.7%)	3(23.0%)	

Data showed mean±SD; proportion(n%); **P-value*<0.05; BMI=Body mass index; AST=Aspartate transaminase; ALT=Alanine aminotransferase; ALP =Alkaline phosphatase; AFP=Alpha fetoprotein; BCLC stage=Barcelona clinic liver cancer stage

4.2 Gut microbial diversity in HCC

From 16S rRNA sequencing preprocessing with FastQC tools to check quality in sequenced reads and Cutadapt to trim primer and adapter from sequencing reads, an average of 37,309.6 ASVs per sample were obtained (Table 3).

Table 3 Preprocessing summary

SampleID	Group	Raw_data	Trimmed_seq	denoisedF	denoisedR	reads_merging(F&R)	input_tax_filter	filtered_tax_filter	percent_filtered_tax
H33	Healthy	145790	108693	104614	107297	91863	44128	44128	100.00
H34	Healthy	111492	78128	75319	76911	65730	36656	36656	100.00
H36	Healthy	121534	85788	84134	85004	76248	45117	45117	100.00
H37	Healthy	102273	74483	70488	73090	59772	32782	32782	100.00
H38	Healthy	114708	82095	79820	81374	75237	58715	58710	99.99
H39	Healthy	102325	73341	69154	72054	60286	34633	34633	100.00
H40	Healthy	117640	82177	77643	80872	68364	47113	47113	100.00
H45	Healthy	131293	94205	91056	92775	76758	33152	33152	100.00
H46	Healthy	115519	86091	84389	85418	77594	54745	54745	100.00
H47	Healthy	113092	80314	76487	78854	63545	32597	32597	100.00
C22	HBV	110706	80014	77551	78848	68643	33351	33351	100.00
C24	HBV	96585	68272	64411	66847	52941	24577	24577	100.00
C32	HBV	140635	105845	100938	104281	87568	54230	54230	100.00
C36	HBV	76760	50597	49393	49941	44202	23944	23944	100.00
C42	HBV	77537	56953	55747	56492	50337	29949	29949	100.00
C56	HBV	86227	62614	61825	62293	59114	44554	44554	100.00
C5	HBV	47038	39133	36200	38032	30712	21481	21481	100.00
C1	HCV	52208	43255	39690	41782	33210	22870	22870	100.00

SampleID	Group	Raw_data	Trimmed_seq	denoisedF	denoisedR	reads_merging(F&R)	input_tax_filter	filtered_tax_filter	percent_filtered_tax
C16	HCV	57361	47399	42048	45316	34975	22336	22336	100.00
C19	HCV	47806	39625	38549	39057	35342	24796	24796	100.00
C20	HCV	49543	39895	38372	39181	33792	24735	24735	100.00
C26	HCV	103710	73481	70331	72398	60977	36327	36327	100.00
C28	HCV	105464	74565	72084	73568	62216	32095	32095	100.00
C34	HCV	101650	72264	71328	71858	66800	41252	41252	100.00
C40	HCV	106415	76362	74510	75378	67565	46899	46899	100.00
C41	HCV	91944	67295	64834	66496	58778	48386	48386	100.00
C60	HCV	103050	78057	76823	77777	73830	71007	71007	100.00
C14	NBNC	50525	42483	41091	41766	36869	27259	27259	100.00
C15	NBNC	50438	42492	40533	41657	36283	25958	25958	100.00
C18	NBNC	54101	45340	43507	44519	39474	27520	27520	100.00
C2	NBNC	53180	43647	41215	42448	36250	24231	24231	100.00
C21	NBNC	105621	78926	77734	78453	73698	58296	58296	100.00
C33	NBNC	108599	74011	69498	72371	58299	34074	34074	100.00
C35	NBNC	86969	64510	63373	64035	59660	45198	45198	100.00
C46	NBNC	72197	52761	51640	52283	47190	33843	33843	100.00
C49	NBNC	113427	75556	72362	74258	60671	41520	41520	100.00
C55	NBNC	111111	75154	68876	73164	58541	41282	41282	100.00
C58	NBNC	103894	73797	71270	72611	61696	36243	36243	100.00
C61	NBNC	106493	77864	75691	77020	69064	49175	49175	100.00
C9	NBNC	45396	38327	36733	37610	32987	25363	25363	100.00

The alpha diversity of species in each sample were significantly decreased in patient with HCC group (P-value < 0.05) (Figure 10A-C). However, there was no difference between viral-related HCC and non-viral-related HCC groups (Figure 11A-C). The beta diversity was calculated with Bray-Curtis and Jaccard distance by NMDS plot, gut microbiome composition in patient with subgroup of HCC and healthy control was significantly separated into two different enterotypes (P = 0.038, Figure 12A-B).

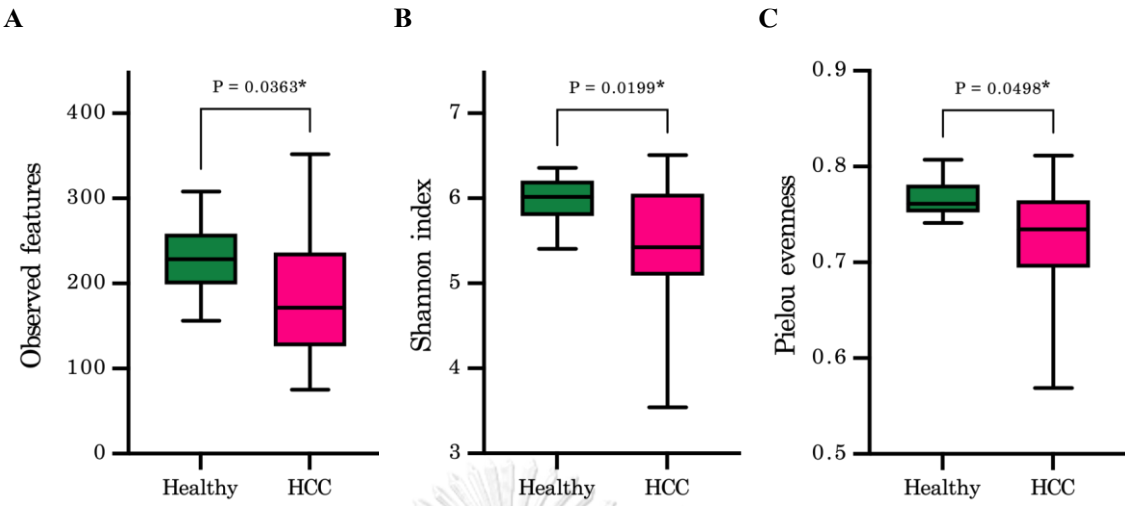


Figure 10 Gut microbiome diversity between healthy and HCC groups **(A)** Alpha diversity; Observed feature **(B)** Shannon index **(C)** Pielou evenness, were significantly decreased in patient HCC (*P = 0.036, 0.020 and 0.050 respectively).

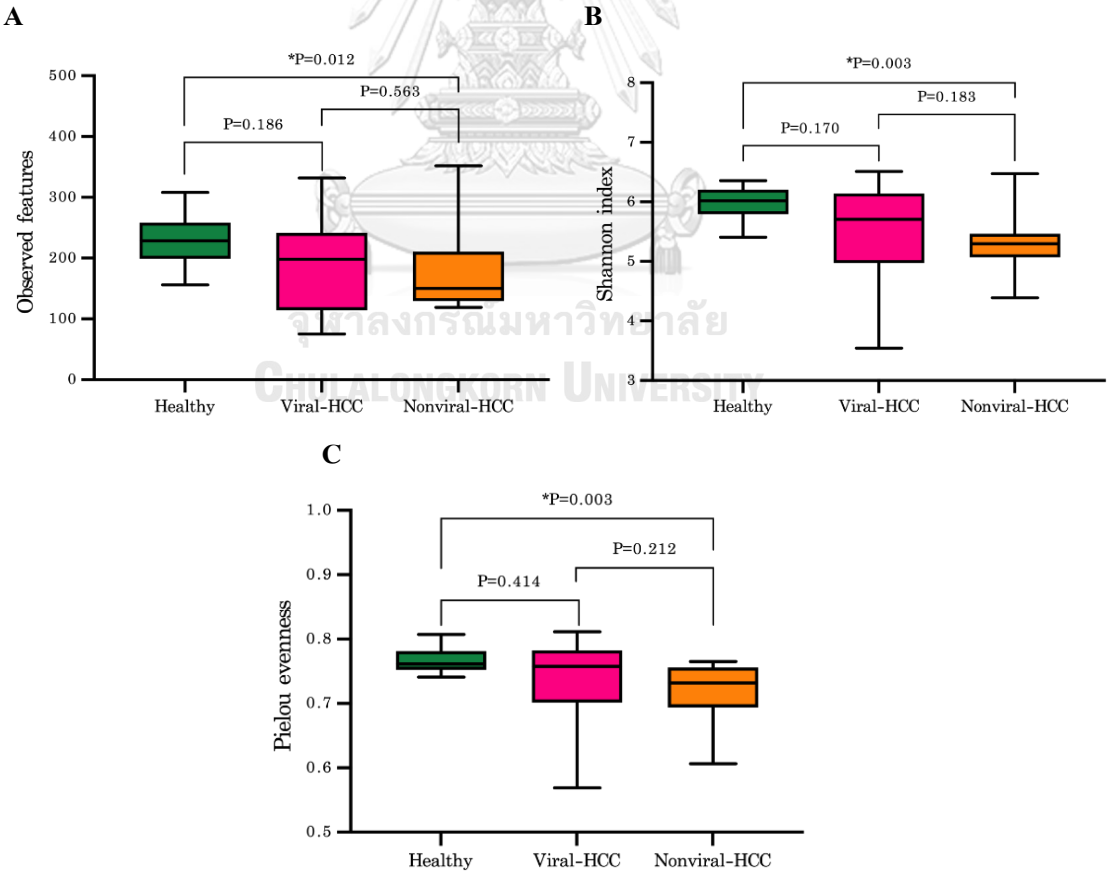


Figure 11 Gut microbiome diversity of all groups **(A)** Alpha diversity; Observed feature **(B)** Shannon index **(C)** Pielou evenness, were significantly decreased in patient with non-viral-related HCC (*P = 0.012, 0.003 and 0.003 respectively). **(D)** Beta diversity; Bray-Curtis distance (*P= 0.038).

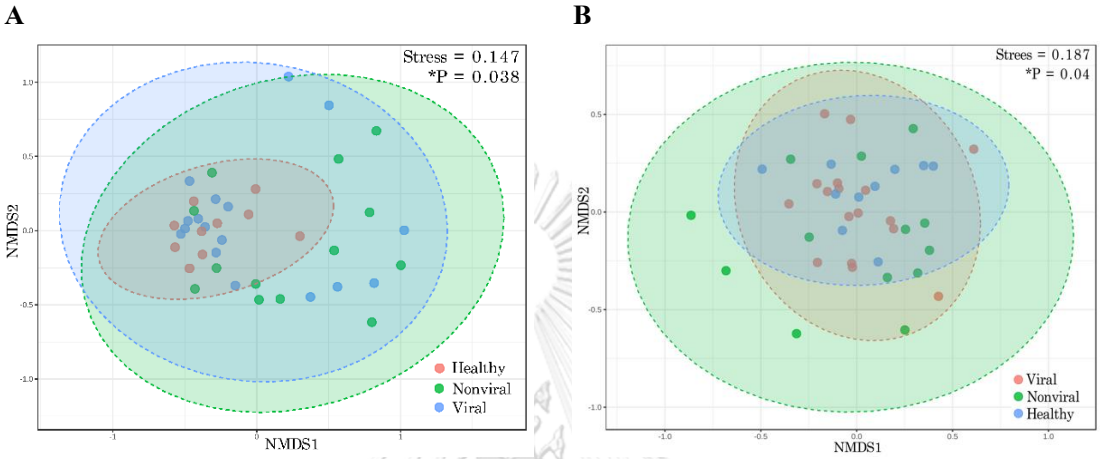
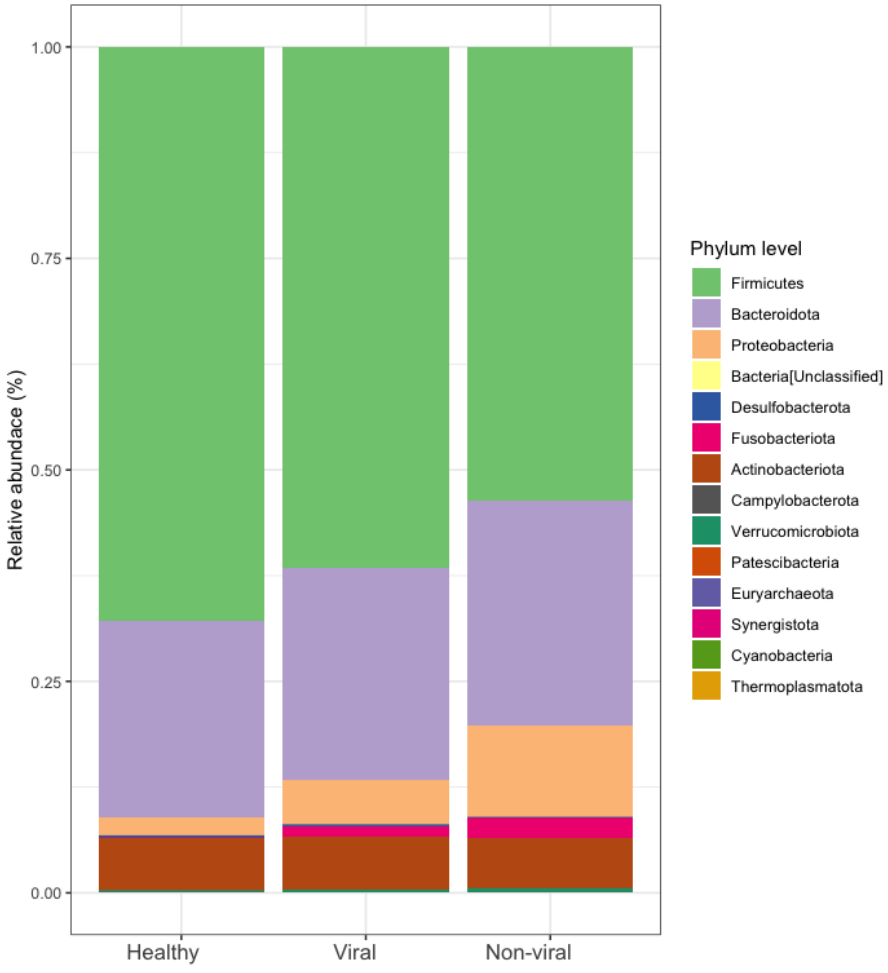


Figure 12 Gut microbiome diversity of all groups **(A)** Beta diversity; Bray-Curtis distance (*P= 0.038). **(B)** Jaccard distance (*P= 0.040)

4.3 Alteration in the composition of gut microbiome associated with HCC

Firmicutes, Bacteroidetes, and Proteobacteria constituted for the majority of the three bacterial phyla in each group on average up to 80% of the ASVs. However, comparison of the most abundances ASVs at phylum showed Proteobacteria significantly increased in HCC group comparison with healthy group (Figure 13A). Regarding the top 50 bacterial genera in terms of relative abundance, it was also evident that 9 genera exhibited significant variations among the different subgroups of HCC (Figure 13B).

A



B

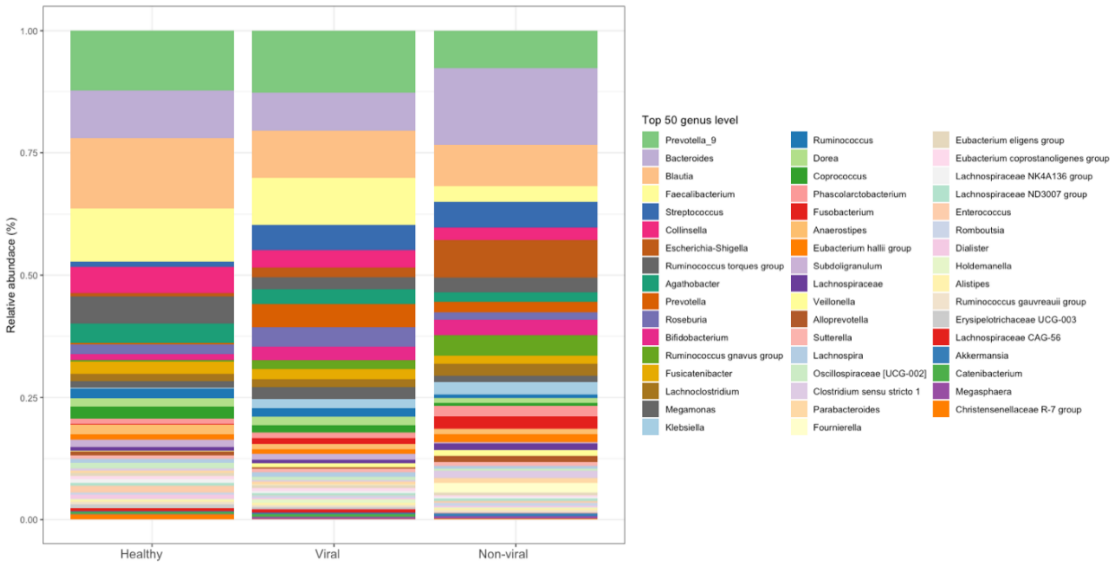


Figure 13 Gut microbiome composition of all participants **(A)** Compositions of gut microbiome at the phylum level between healthy controls and HCC subgroups. **(B)** Compositions of gut microbiome at the top 50 genus level between healthy controls and HCC subgroups.

PICRUSt2 was utilized to predict the functional analysis of microbial communities in different subgroups of HCC. In participants with non-viral-related HCC, the mean proportion increased and significantly predominant including lipopolysaccharide biosynthesis, fatty acid metabolism and dioxin degradation (Figure 14).

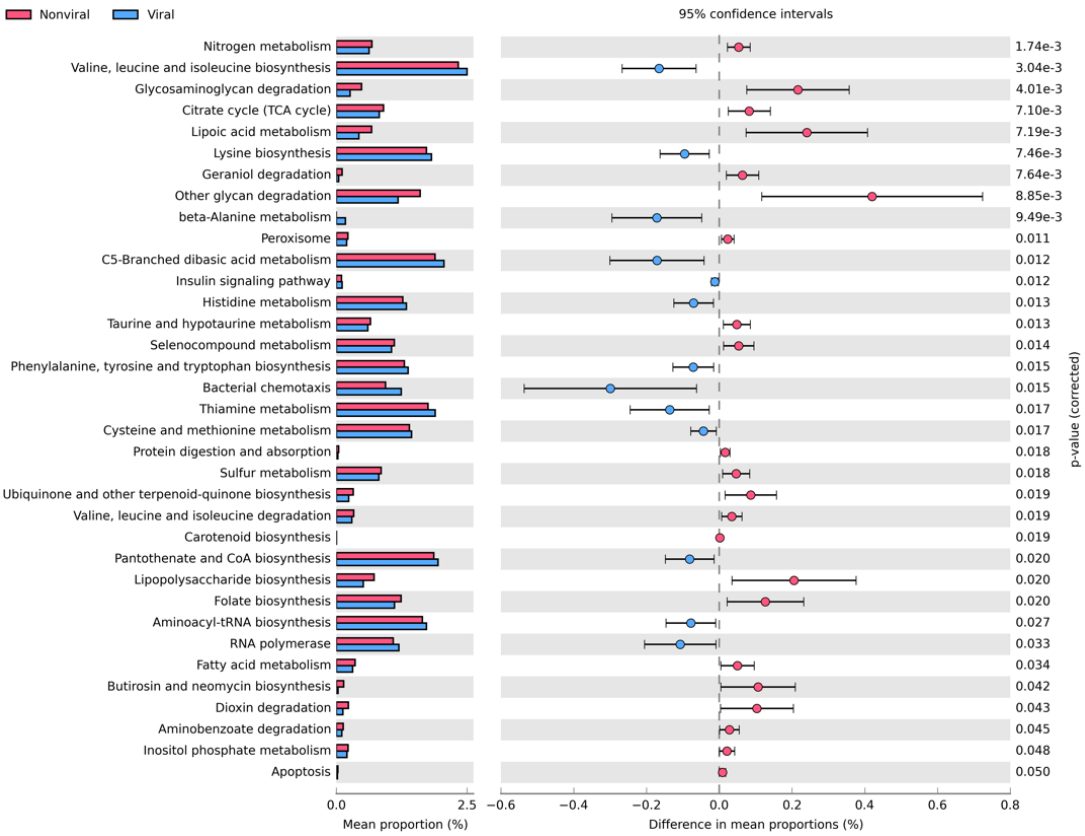
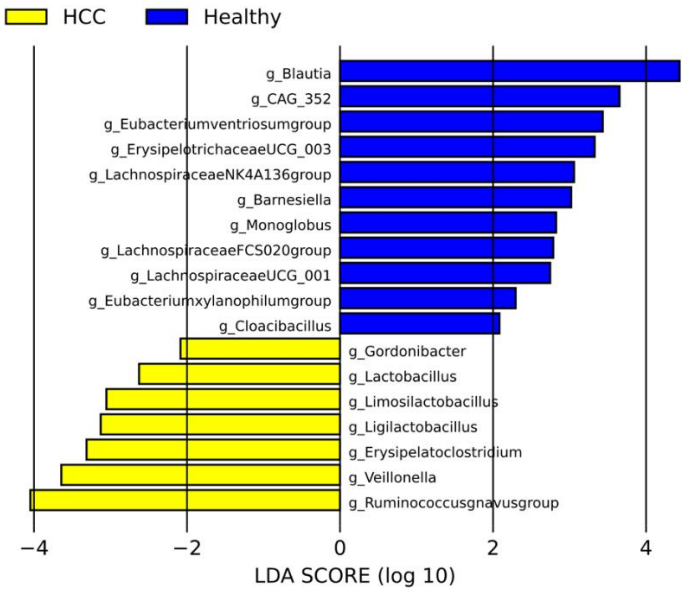


Figure 14 Functional pathways predicted by PICRUSt2 that differentiate in viral-related HCC and non-viral-related HCC.

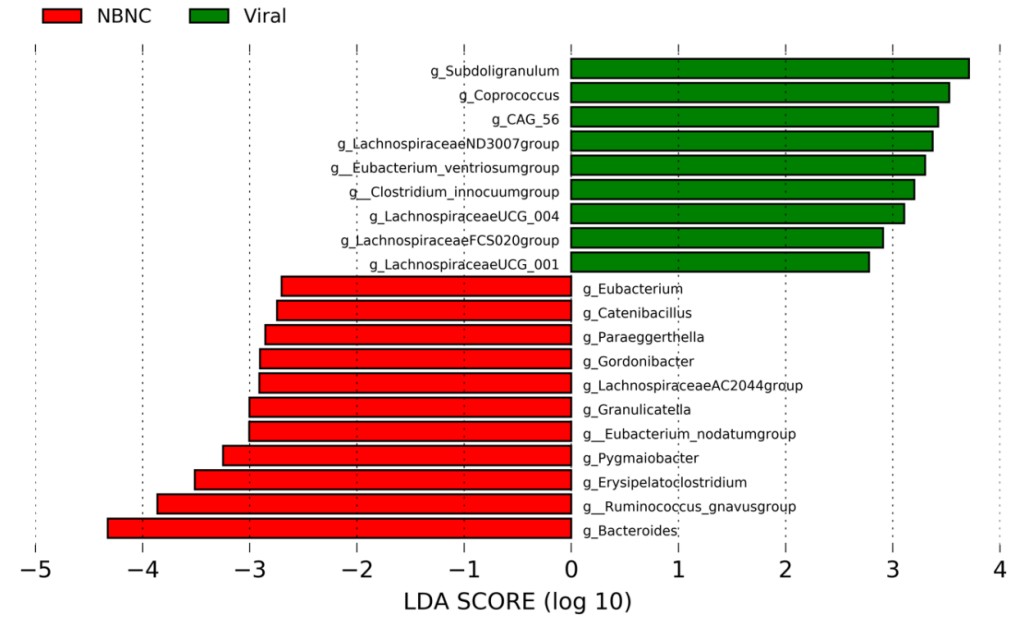
LEfSe was utilized to identify bacterial taxa linked to healthy, HCC and various causes within the HCC group. Total 18 bacterial taxa differences in microbiota compositions between heathy and HCC (Figure 15A). Moreover, 11 bacterial taxa including *Eubacterium*,

Catenibacillus, *Paraeggerthella*, *Gordonibacter*, *Lachnospiraceae AC2044 group*, *Granulicatella*, *Eubacterium nodatum group*, *Pygmaiobacter*, *Erysipelatoclostridium*, *Ruminococcus gnavus group* and *Bacteroides* exhibited significant overrepresentation with a $\log_{10}$ LDA score > 2 in the fecal samples of patients belonging to the non-viral-related HCC subgroup (Figure 15B-C).

A



B



C

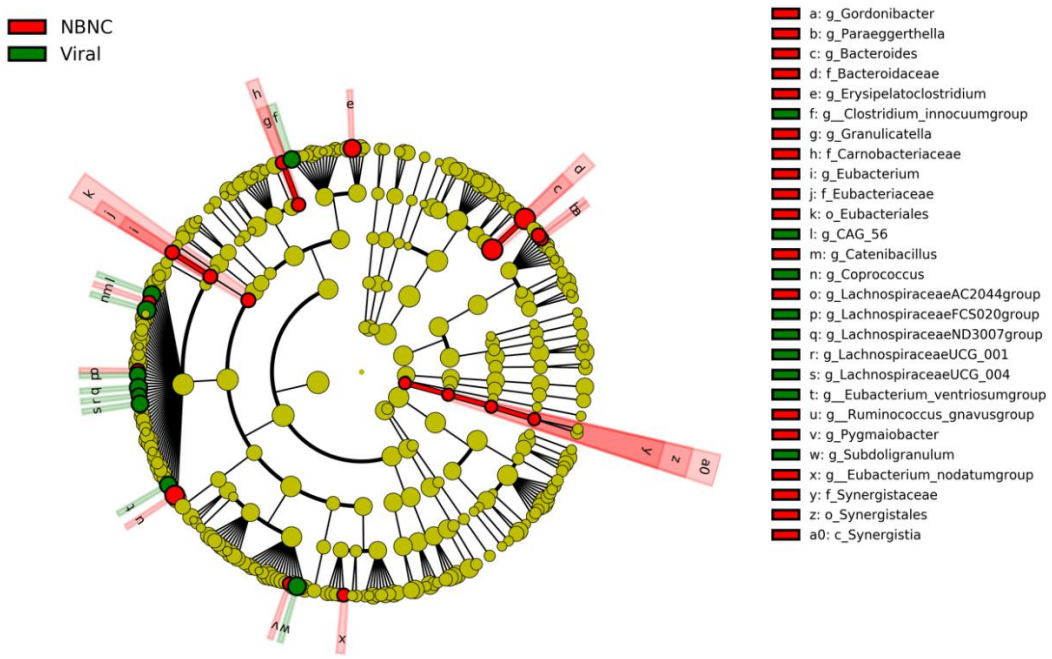


Figure 15 LEfSe analysis of differential gut microbial in genus level. **(A)** Histograms LDA score between healthy and HCC group. **(B)** Histograms LDA score between HCC subgroups. **(C)** Cladogram between HCC subgroups.

4.4 Overview of host transcriptome in subgroup of HCC

We hypothesized that change in host transcriptome may be correlated with change in gut microbiome. Thus, we performed a transcriptome analysis of total RNA expression profile from PBMCs between healthy and HCC group, a total of 261 genes were identified to be differentially expressed in HCC patients, which included 39 up-regulated (P-value < 0.05, log2FC > 1.5) and 222 down-regulated genes (P-value < 0.01, log2FC < -1.5) in HCC group (Figure 16A) Moreover, we performed a transcriptome analysis of total RNA expression profile from PBMCs of 17 patients with viral-related HCC and 13 patients with non-viral-related HCC, a total of 80 genes were identified to be differentially expressed in subgroup of HCC patients, which included 70 up-regulated (P-value < 0.05, log2FC > 1.5) and 10 down-regulated genes (P-value < 0.01, log2FC < -1.5) in non-viral-related HCC (Figure 16B).

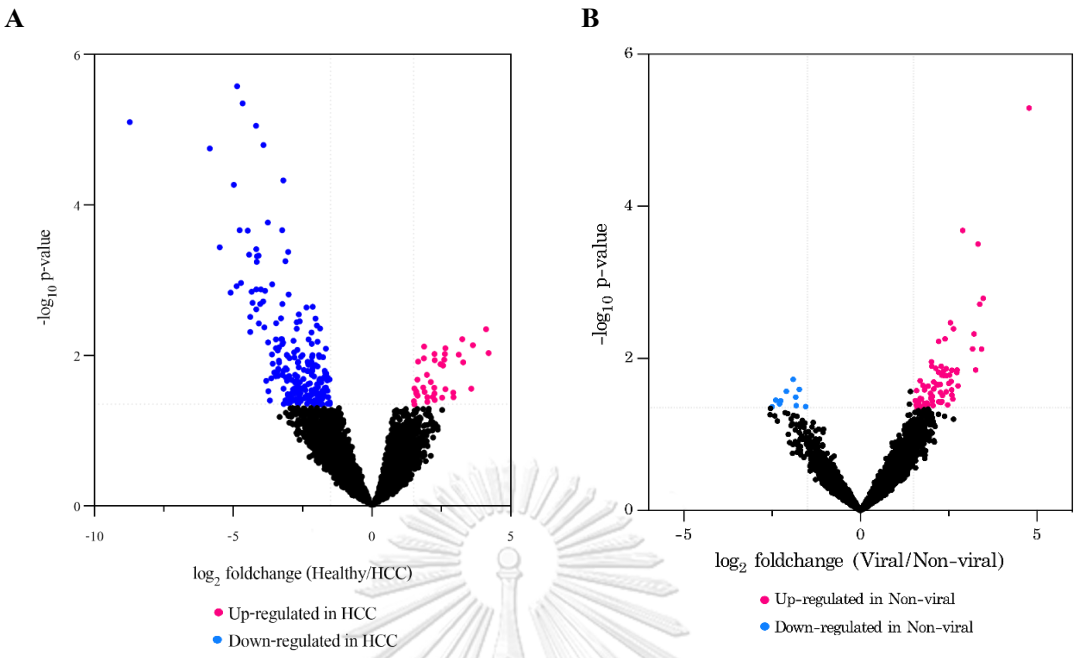


Figure 16 Transcriptome profiles of PBMCs. **(A)** A volcano plot of differential gene expression of healthy and HCC group. **(B)** A volcano plot of differential gene expression of non-viral-related HCC compared with viral-related HCC.

4.5 Association of host transcriptome profile influenced by gut microbiome

Based on Pearson's correlation analyses, we tested the associations between host gene expressed ang gut microbiome to discover host-gut microbe and to clarify how gut microbiota influence the transcriptome profiles of HCC. A total of 6,137 genes and 268 gut microbes were performed. A total of 1,644,716 gene pairs were calculated. Total of 23 genes and 7 gut microbes were identified as positively correlated (Table 4).

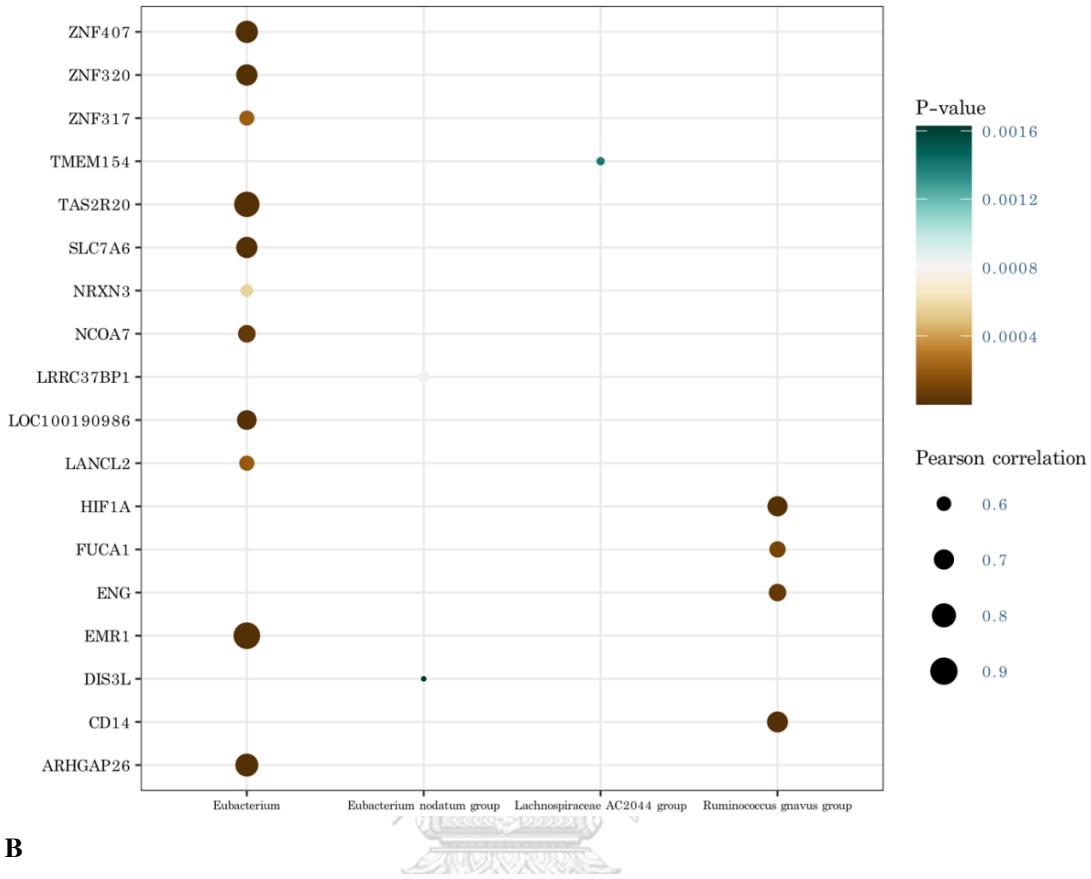
Table 4 Twenty-four gut-gene pairs filtered by Pearson's coefficient correlation

Gene symbol	Gut microbiome	Pearson's coefficient	P-value	FDR
EMR1	*Eubacterium*	0.92480526	2.84E-13	6.17E-10
TAS2R20	*Eubacterium*	0.881899	1.20E-10	1.23E-07
FRMD3	*Eubacterium ventriosum group*	0.81688016	3.67E-08	1.60E-05
ARHGAP26	*Eubacterium*	0.80716509	7.09E-08	2.75E-05
KIFC3	*Eubacterium ventriosum group*	0.80376367	8.85E-08	3.29E-05
ZNF407	*Eubacterium*	0.78740772	2.44E-07	7.58E-05
ZNF320	*Eubacterium*	0.761265	1.04E-06	0.00025026

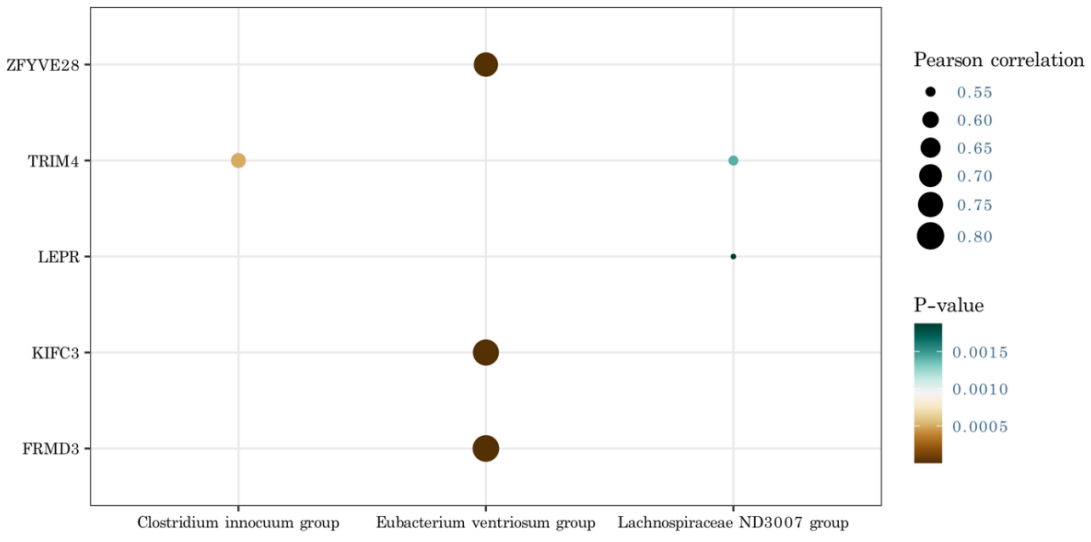
Gene symbol	Gut microbiome	Pearson's coefficient	P-value	FDR
ZFYVE28	*Eubacterium ventriosum group*	0.76107347	1.05E-06	0.00025221
SLC7A6	*Eubacterium*	0.7600947	1.10E-06	0.00026271
CD14	*Ruminococcus gnavus group*	0.7516247	1.69E-06	0.00037421
HIF1A	*Ruminococcus gnavus group*	0.73081233	4.52E-06	0.00084218
LOC100190986	*Eubacterium*	0.71927356	7.51E-06	0.00127084
NCOA7	*Eubacterium*	0.67244723	4.69E-05	0.00559694
ENG	*Ruminococcus gnavus group*	0.67139189	4.87E-05	0.00576831
FUCA1	*Ruminococcus gnavus group*	0.64891485	0.00010489	0.01064955
LANCL2	*Eubacterium*	0.62938557	0.00019449	0.01735448
ZNF317	*Eubacterium*	0.62737199	0.00020679	0.01819793
TRIM4	*Clostridium innocuum group*	0.59681349	0.00049896	0.03606712
NRXN3	*Eubacterium*	0.59242548	0.00056217	0.03945515
LRRC37BP1	*Eubacterium nodatum group*	0.57719875	0.00083953	0.05349469
TMEM154	*Lachnospiraceae AC2044 group*	0.55665891	0.00139951	0.07875074
TRIM4	*Lachnospiraceae ND3007 group*	0.55603137	0.00142081	0.07966113
DIS3L	*Eubacterium nodatum group*	0.55034436	0.001627	0.0882571
LEPR	*Lachnospiraceae ND3007 group*	0.54426299	0.0018758	0.09800949

The increased relative abundance of gut microbe (*Eubacterium, Eubacterium nodatum group, Lachnospiraceae AC2044 group* and *Ruminococcus gnavus group*) were associated with significant up regulated genes in patients with non-viral-related HCC (Figure 17A). Functional and pathway analysis exposed set of gene to play an important role in the identification of signal transduction, programmed cell death, neuronal system, metabolism of proteins, immune system and disease. Based on Reactome pathway database, the host-gut microbes were significantly about disease and immune system, which involved immune regulation (Figure 17C). Moreover, the decreased abundance of gut microbe (*Clostridium innocuum group, Eubacterium ventriosum group* and *Lachnospiraceae ND3007 group*) were associated with significant down regulated genes in patients with non-viral-related HCC (Figure 17B).

A


ZNF407
ZNF320
ZNF317
TMEM154
TAS2R20
SLC7A6
NRXN3
NCOA7
LRRC37BP1
LOC100190986
LANCL2
HIF1A
FUCA1
ENG
EMR1
DIS3L
CD14
ARHGAP26
Eubacterium
Eubacterium nodatum group
Lachnospiraceae AC2044 group
Ruminococcus gnavus group
P-value
0.0016
0.0012
0.0008
0.0004
Pearson correlation
0.6
0.7
0.8
0.9

B


ZFYVE28
TRIM4
LEPR
KIFC3
FRMD3
Clostridium innocuum group
Eubacterium ventriosum group
Lachnospiraceae ND3007 group
Pearson correlation
0.55
0.60
0.65
0.70
0.75
0.80
P-value
0.0015
0.0010
0.0005

C

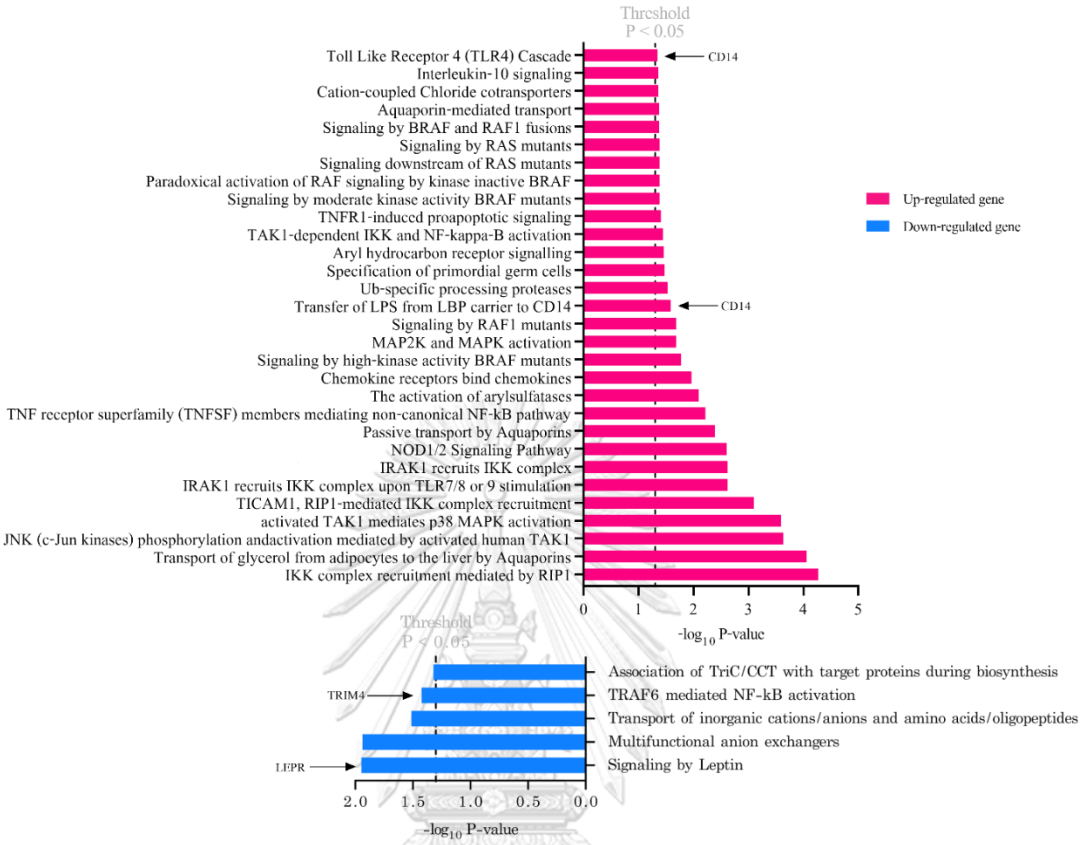


Figure 17 The association between gut microbiome and host transcriptome in HCC subgroups. **(A)** Pearson's correlation with up regulated host genes and increased gut microbe in non-viral-related HCC group. **(B)** Pearson's correlation with down regulated host genes and decreased gut microbe in viral-related HCC group. **(C)** Functional analysis of differentially regulated genes between patients with HCC subgroups. Regarding up-regulated genes, the immune response and inflammatory pathways involving the pro-inflammatory genes are among the most significantly enriched pathways. The dashed line indicates the Fisher exact test P value threshold set at 0.05.

To clarify the localization and functions of these genes within the blood-immune microenvironment, we examined their expressions in a separate single-cell mapping database (using SMART-seq2) of hepatocellular carcinoma (HCC) [79]. Interestingly, genes associated with gut microbes were high expressed in B cell, macrophages, $CD8^+$ T cells, $CD4^+$ T cells and NK cells (Figure 18). The genes exhibited an interconnected relationship, indicating that the gut

microbiota potentially influences the transcriptome of hepatocellular carcinoma (HCC) through various factors.

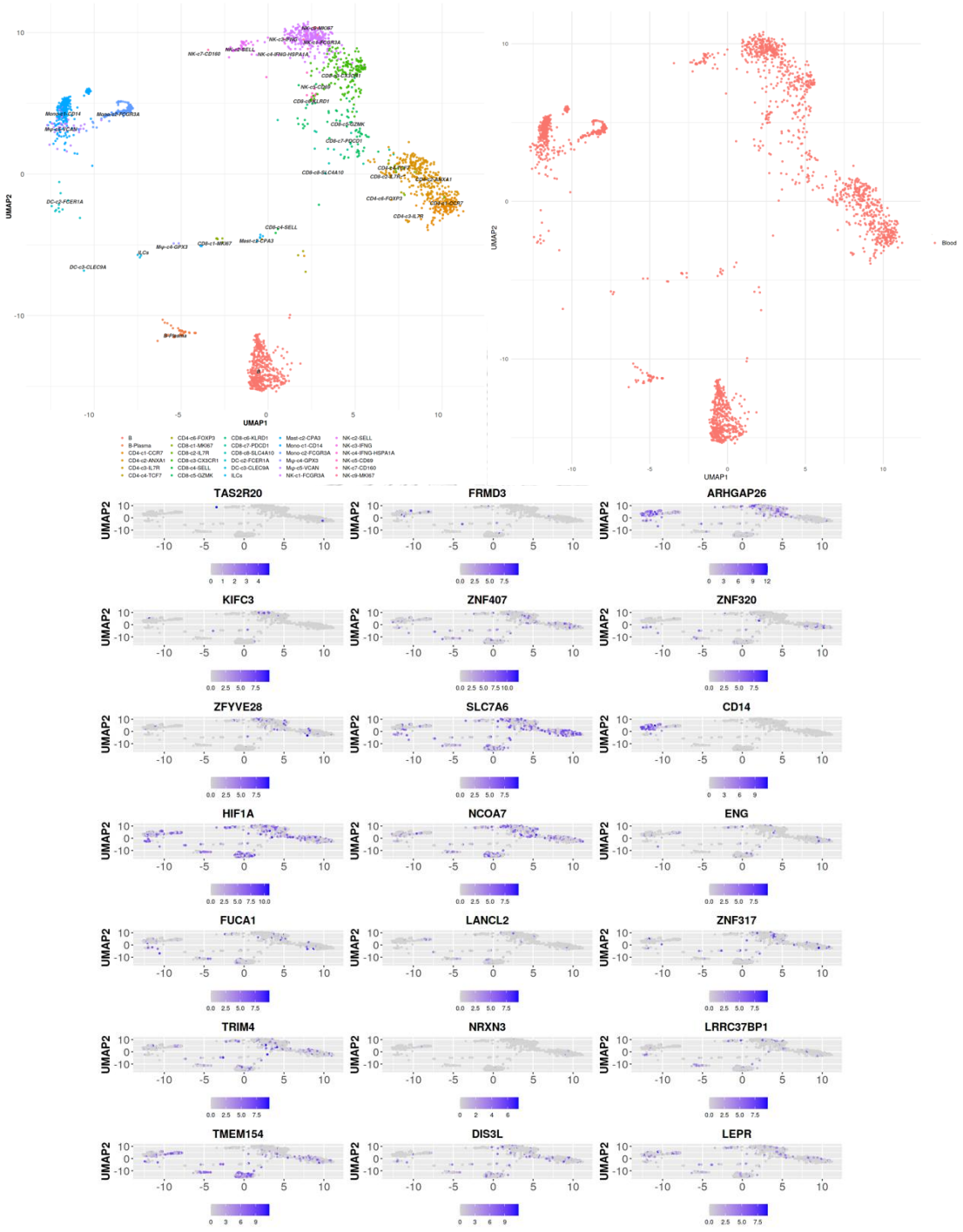


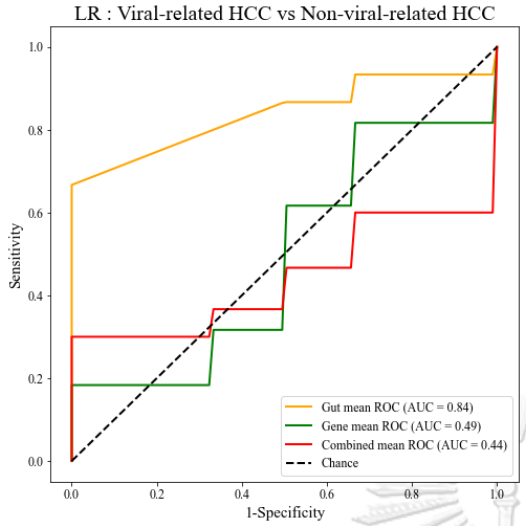
Figure 18 The expressions of host genes related to the gut were examined for each cell type using SMART-seq2 data (http://cancer-pku.cn:3838/HCC). Uniform Manifold Approximation and Projection (UMAP) plots were generated to visualize the cell clusters identified through

integrated analysis, with each cluster represented by a distinct color (first plot). UMAP plots depict the distribution of cells across sample types (second plot). UMAP plots depict the distribution of cells for each specific gene (third plot).

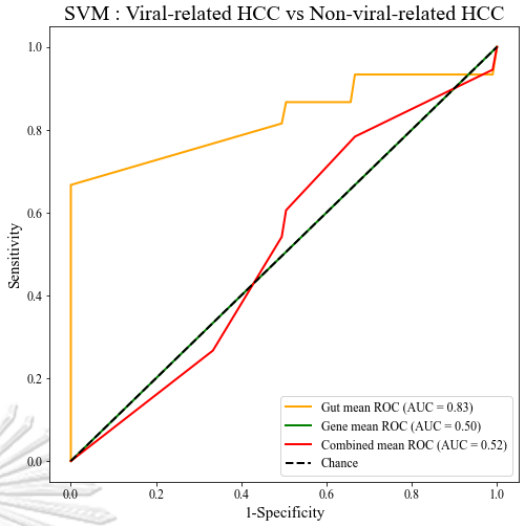
4.6 Gut microbiome and gene marker for HCC subgroups classification

In this study, we explored the distinct characteristics of commonly used machine learning algorithms for the analysis of multi-omics data, emphasizing the critical importance of algorithm selection. Our investigation focused on classifying HCC subgroups, employing three prominent ML algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The predictive performance, assessed through fivefold cross-validation, revealed compelling results. For gene expression data, a set of 18 genes exhibited significant positive correlations among patients with non-viral-related HCC, demonstrating strong diagnostic potential for HCC. These genes included EMR1, TAS2R20, ARHGAP26, ZNF407, ZNF320, SLC7A6, LOC100190986, NCOA7, LANCL2, ZNF317, NRXN3, CD14, HIF1A, ENG, FUCA1, TMEM154, LRRC37BP1, and DIS3L. In parallel, gut microbiome data featured four genera (*Eubacterium*, *Eubacterium nodatum group*, *Lachnospiraceae AC2044 group*, and *Ruminococcus gnavus group*) that exhibited similarly positive correlations in non-viral-related HCC patients. When evaluating classification performance, the integrated mean Area Under the Curve (AUC) values underscored the robustness of LR (0.84), SVM (0.83), and RF (0.82) for gut microbiome variables (Figure 19A-D). Notably, combining both gene expression and gut microbiome data did not yield optimal classification results for HCC subgroups. However, leveraging the Synthetic Minority Over-sampling Technique (SMOTE) to address dataset imbalance, particularly with 17 viral-related HCC and 17 non-viral-related HCC cases, enhanced the Random Forest algorithm's mean AUC to 0.85 (Figure 20A-D). Remarkably, the combined dataset achieved the highest mean AUC values, reaching 0.87 (Figure 20C). These findings underscore the potential of specific gut microbiome markers in elucidating disease causation and their promising role in distinguishing between patients with viral-related and non-viral-related HCC. The integration of multi-omics data and strategic algorithm selection emerges as a powerful strategy for advancing HCC subgroup classification and enhancing diagnostic accuracy.

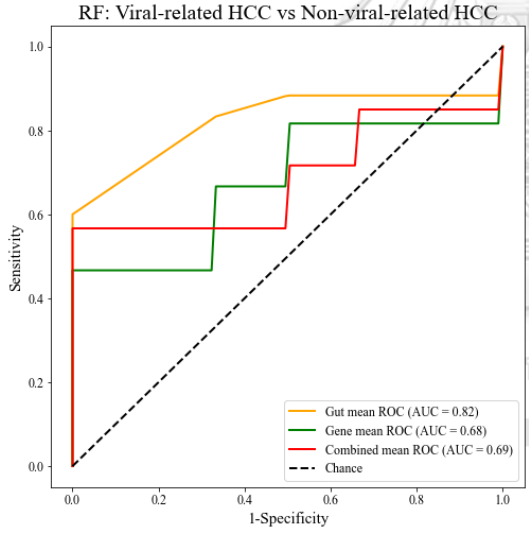
A



B



C

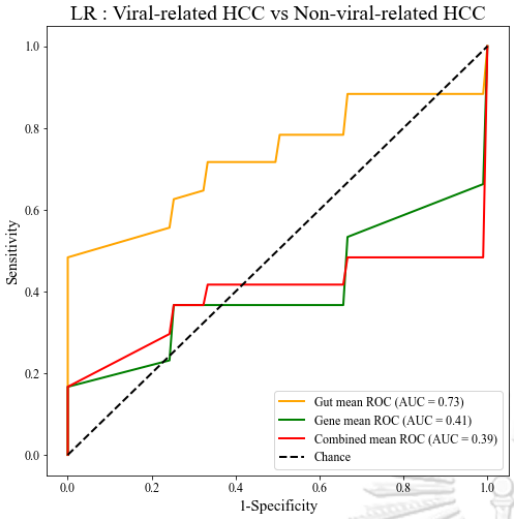


D

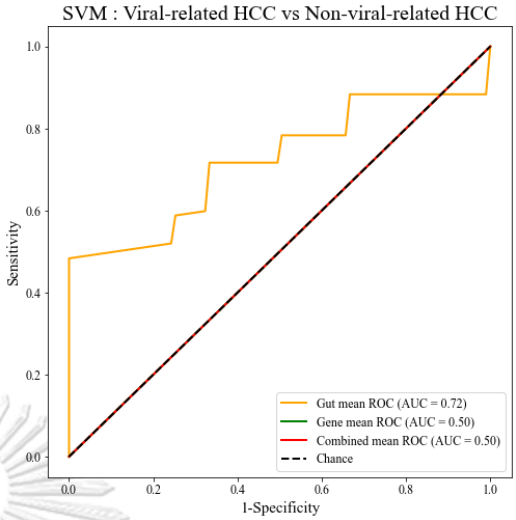
Model	Biomarker	Mean sensitivity (%)	Mean specificity (%)	Mean accuracy (%)	Mean AUC
Logistic Regression	Gut	80	20	57	0.84
	Gene	70	40	57	0.49
	Gut+Gene	47	23	37	0.44
Support Vector Machine	Gut	100	0	57	0.83
	Gene	100	0	57	0.5
	Gut+Gene	100	0	57	0.52
Random Forest	Gut	68	43	60	0.82
	Gene	82	53	70	0.68
	Gut+Gene	75	63	67	0.69

Figure 19 Receiver operating characteristic analysis of our model classification of non-viral related HCC versus viral related HCC. The true-positive rate (sensitivity) is plotted against the false-positive rate (1-specificity). The mean AUC values of ROC curves with fivefold cross-validations are gene expression set, gut microbiome set and combined of two datasets for classification model. **(A)** Logistic Regression model. **(B)** Support Vector Machine model. **(C)** Random forest model. **(D)** Summary of evaluation matrix.

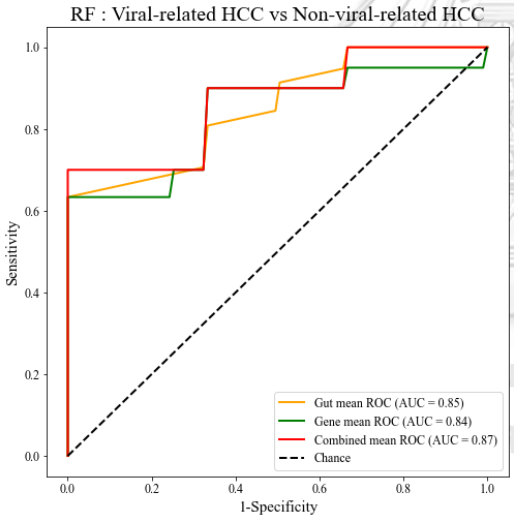
A



B



C



D

Model	Biomarker	Mean sensitivity (%)	Mean specificity (%)	Mean accuracy (%)	Mean AUC
Logistic Regression	Gut	82	47	65	0.85
	Gene	42	65	53	0.84
	Gut+Gene	42	72	58	0.87
Support Vector Machine	Gut	53	40	41	0.72
	Gene	60	40	44	0.5
	Gut+Gene	60	40	44	0.5
Random Forest	Gut	88	80	86	0.85
	Gene	63	95	77	0.84
	Gut+Gene	80	80	83	0.87

Figure 20 Receiver operating characteristic analysis of our model classification of non-viral related HCC versus viral related HCC after using SMOTE technique. The true-positive rate (sensitivity) is plotted against the false-positive rate (1-specificity). The mean AUC values of ROC curves with fivefold cross-validations are gene expression set, gut microbiome set and combined of two datasets for classification model. **(A)** Logistic Regression model. **(B)** Support Vector Machine model. **(C)** Random Forest model. **(D)** Summary of evaluation matrix.

Chapter 5 Discussion and conclusion

The previous report [63] and our study have shown that the gut community is highly different in any cohort study, which is associated with host transcriptome. Recent findings have demonstrated that the progression of HCC is caused by genetic and epigenetic changes acquired via repetitive hepatocyte destruction and regeneration. Our study is the first to report an association between two datasets from any causation of HCC including viral-related HCC and non-B-non-C (NBNC) or non-viral-related HCC. The diversity of fecal microbiota was found to be significantly lower in HCC groups compared to the healthy control group. However, no significant difference in fecal microbial diversity was observed between the non-viral-related HCC group and the viral-related HCC group. In general, patients with viral-related HCC exhibited greater species richness. In non-viral-related HCC patients, there was a decrease in the abundance of Firmicutes and an increase in Proteobacteria at the phylum level. Our findings revealed a distinct pattern in the gut microbiota composition between non-viral-related HCC and viral-related HCC patients. Specifically, non-viral-related HCC patients exhibited a decrease in potential anti-inflammatory bacteria and an increase in pro-inflammatory bacteria. Conversely, viral-related HCC patients demonstrated a higher abundance of potential anti-inflammatory bacteria. These results suggest that the gut microbiota may have a significant impact on the progression of viral or non-viral-related HCC. The evidence showed associated with the presence of particular gut microbes. Distinguished by the enrichment *Eubacterium, Catenibacillus, Paraeggerthella, Gordonibacter, Lachnospiraceae AC2044 group, Granulicatella, Eubacterium nodatum group, Pygmaiobacter, Erysipelatoclostridium, Ruminococcus gnavus group and Bacteroides* are significantly enriched in non-viral-related HCC group, while *Subdoligranulum*, *Coprococcus*, *CAG_56*, *Lachnospiraceae ND3007 group*, *Eubacterium ventriosum group*, *Clostridium innocuum group*, *Lachnospiraceae UCG_004*, *Lachnospiraceae FCS020 group* and *Lachnospiraceae UCG_001* were significantly increased in viral-related HCC group. In another study of exploring what features of gut microbiota are associated with cirrhosis hepatocellular carcinoma (HCC) and non-alcoholic fatty liver disease (NAFLD), the results showed high abundance levels of *Bacteroides* and *Ruminococcaceae* suggested that gut microbiota are significantly correlated with systemic inflammation in the process of hepatocarcinogenesis [59].

Differences in the composition of gut bacteria between non-B-non-C causes may play a role in disease development through various pathways [80]. For example, an increase in the expression levels of specific pro-inflammatory cytokines in the liver. Research findings indicate that a decline in microbial diversity is linked to an increase in intestinal permeability and the presence of systemic low-grade inflammation [81]. As a consequence, this association has been connected to the development of hepatic steatosis. As for subgroups of non-viral-related HCC specific gut microbial signatures, which correlated with host gene transcriptome. We observed that patients with 4 genera, *Eubacterium, Eubacterium nodatum group, Lachnospiraceae AC2044 group* and *Ruminococcus gnavus group* are positively correlated with several of the disease and immune system. *Eubacterium* is a bacterium that is classified as an obligate anaerobe and utilizes dietary fiber through fermentation to generate short-chain fatty acids (SCFAs), which include butyric acid [82]. The involvement of SCFAs in the pathogenesis of NAFLD is crucial as they have the potential to influence and maintain intestinal homeostasis, while also positively affecting glucose and lipid metabolism. In human peripheral blood mononuclear cells (PBMC), SCFAs such as propionate and butyrate have been found to suppress the expression of lipopolysaccharide (LPS)-induced cytokines, specifically interleukin-6 (IL-6) and IL-12p40. The liver functions as a source of inflammatory agents and plays a pivotal role in mounting inflammatory responses to bacterial endotoxins, also known as lipopolysaccharide (LPS) [83]. Kupffer cells (KCs) are the specialized macrophages naturally present in the liver. Their main role involves removing bacteria and soluble bacterial byproducts, while also producing inflammatory cytokines [84]. Toll-like receptor 4 (TLR4) is a type of pattern-recognition receptor (PRR) found on the surface of Kupffer cells. Its primary function is to detect the presence of microbes and LPS [85]. Several studies have indicated that short-chain fatty acids (SCFAs), including butyrate, can contribute to the development of colorectal cancer by promoting the conversion of colonic epithelial cells and causing abnormal cell growth [86]. In our study, the gut microbiome includes several SCFAs-producing bacteria that are associated with host genes. Moreover, the prediction of pathway analysis via PICRUSt2 showed differences between the two subgroups of HCC including lipopolysaccharide biosynthesis that might be stimulate cytokine release, demonstrating TLR4 selectivity in recognition.

Additionally, we found that *Ruminococcus gnavus group* was associated with up regulated CD14 of non-viral-related HCC patients. The abundance of *Ruminococcus gnavus group* increases in patients with liver disease, especially hepatocellular carcinoma [87]. A previous study demonstrated that the *Ruminococcus gnavus group* generates glucorhamnan, which acts as a TLR4 ligand, leading to the subsequent release of tumor necrosis factor-alpha (TNF-α) by dendritic cells [88]. Notably, our study showed *Ruminococcus gnavus group* exhibited the most positive association to CD14 and contributed to liver inflammation. The translocated LPS is recognized by CD14 and TLR4, triggering the release of pro-inflammatory cytokines like tumor necrosis factor alpha (TNFα), interferon alpha (IFNα), interferon-gamma (INFγ), and interleukins (IL1β or IL6). This can ultimately lead to the onset of endotoxemia [89]. The functional analysis of the gut microbiome in our findings indicates a connection with endotoxin and inflammation, which were influenced by various subgroups of HCC. Among them, lipopolysaccharide biosynthesis were identified exactly in non-viral-related HCC group from viral-related HCC group. LPS can be released from the outer membrane during gram negative bacterial growth, death, or antibiotic treatment. This release depends on factors like bacterial death type, antibiotic concentration, and incubation conditions [90]. Most immune cells express TLR4 and activate signaling pathways upon LPS binding. CD14, a membrane protein, binds LPS before TLR4 activation and transfers it to Lymphocyte antigen 96 (MD2), a protein complexed with TLR4. This leads to two signaling pathways: Myd88 and Toll interleukin-1 receptor domain-containing adapter-inducing interferon-dependent pathways. These pathways result in the transcription of proinflammatory cytokines such as IL-8, IL-6, IL-1, IL-12, IFN, and TNF [91]. Recently, there has been an established link between the bloom of the Gram-positive bacterium *Ruminococcus gnavus* and the onset of inflammatory bowel disease. Additionally, a recently discovered polysaccharide produced by this bacterium has been demonstrated to induce the release of inflammatory cytokines. It has been hypothesized that this stimulation occurs through the activation of toll-like receptor 4 (TLR4) [92]. Moreover, this study highlights the significance of algorithm selection in analyzing multi-omics data for Hepatocellular Carcinoma (HCC) subgroups. Utilizing Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), we identified 18 genes and four gut microbiome genera with diagnostic potential for non-viral-related HCC. LR, SVM, and RF demonstrated robust classification

performance for gut microbiome variables, with mean AUC values of 0.84, 0.83, and 0.82, respectively. The integration of gene expression and gut microbiome data did not yield optimal results. Nevertheless, employing the Synthetic Minority Over-sampling Technique (SMOTE) proved beneficial, notably enhancing the Random Forest model mean AUC to 0.85. Remarkably, the combined dataset achieved the highest mean AUC of 0.87, underscoring its potential in enhancing diagnostic accuracy and disease subgroup classification.

In conclusion, this study has provided valuable insights into the potential significance of differential gene expression correlated with the gut microbiome in relation to various etiological factors of HCC. It opens up new avenues for exploring human biomarker discovery. Further investigations are required to validate our results and gain a deeper understanding of the mechanisms involved in the interaction between the host-gut microbiome and metabolites in patients with HCC.

Chapter 6 Limitation and suggestion

Our study has several limitations. Firstly, it is a retrospective study without follow-up data, including overall survival (OS) information. As a result, overall survival data is essential for evaluating the ultimate impact of a treatment protocol on a patient's lifespan. Without such data, researchers may only be able to assess short-term outcomes, which might not provide a comprehensive picture of the treatment's efficacy or potential side effects over the long term. This limitation can hinder the ability to make evidence-based recommendations and may necessitate reliance on surrogate endpoints or extrapolation, which can introduce uncertainty and potential bias into the analysis

Secondly, our sample size was relatively small. Due to the difficulty in recruiting the patients following the inclusion and exclusion criteria, our sample collection did not reach the minimum sample size. Small sample sizes can severely limit the generalizability of machine learning models in clinical research. Models may not capture the full range of variability and complexity present in the patient population. As a result, the models may perform well on the limited data they were trained on but struggle to generalize to new, unseen data or different patient populations. Additionally, small sample sizes can lead to reduced statistical power. it becomes challenging to identify statistically significant patterns or make confident conclusions about the effectiveness of a treatment, the presence of rare adverse events, or the accuracy of diagnostic models. This limitation can lead to false positives or false negatives, making it difficult to draw reliable conclusions from the machine learning analysis.

Thirdly, when analyzing the gut microbiome and host transcriptome without including metabolome data, there is a significant gap in our understanding of host-microbiome interactions. Metabolites are the small molecules produced by both the host and the gut microbiota as a result of metabolic processes. They play a crucial role in mediating the crosstalk between the two entities. Metabolites can act as signaling molecules, energy sources, and regulators of various biological processes, and may affect the composition and activities of the microbiome. To overcome this limitation, researchers often strive to obtain metabolome data in addition to microbiome and transcriptome data, enabling a more holistic understanding of host-microbiome interactions and their impact on health and disease. Integrating all three types of data (metabolome, microbiome, and transcriptome) can lead to more comprehensive insights and help unravel the complex mechanisms at play in the gut ecosystem.

REFERENCES

1. WHO. *http://gco.iarc.fr/today/data/factsheets/cancers/11-Liver-fact-sheet.pdf*. Liver 2018.
2. Josep M. Llovet, R.K.K., Augusto Villanueva, Amit G. Singal, Eli Pikarsky, Sasan Roayaie, Riccardo Lencioni, Kazuhiko Koike, Jessica Zucman-Rossi and Richard S. Finn, *Hepatocellular carcinoma.* Nature Review, 2021. **7**.
3. Ju Dong Yang, P.H., Gregory J. Gores, Amina Amadou, Amelie Plymoth and Lewis R. Roberts, *A global view of hepatocellular carcinoma: trends, risk, prevention, and management.* Nature Reviews., 2019. **16**: p. 590–604.
4. Daniela Sciancalepore, M.T.Z., Chiara Valentina Luglio, Carlo Sabbà and Nicola Napoli, *Hepatocellular Carcinoma: Known and Emerging Risk Factors.* Journal of Cancer Therapy, 2018. **9**.
5. M, B.J.a.S., *AASLD Practice Guidelines: management of hepatocellular carcinoma: an update.* Hepatology, 2010: p. 1-35.
6. Asham EH, K.A.a.G.R., *Management of hepatocellular carcinoma.* Surg Clin North Am, 2013. **9**: p. 1423-1450.
7. Kashyap R, J.A., Nalesnik M, Carr B, Barnes J, Vargas HE, Rakela J and Fung J, *Clinical significance of elevated α-fetoprotein in adults and children. Digestive diseases and sciences.* Digestive diseases and sciences, 2001. **46**.
8. PJ, J., *The role of serum alpha-fetoprotein estimation in the diagnosis and management of hepatocellular carcinoma.* Clinics in liver disease, 2001. **5**.
9. Appanna, V.D., *Human Microbes - The Power Within Health, Healing and Beyond.* Spingers, 2017.
10. Peter J. Turnbaugh, R.E.L., Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon, *The Human Microbiome Project.* Nature, 2007. **449**.
11. Di Ciaula A, B.J., Garruti G, Celano G, De Angelis M, Wang HH, Di Palo DM, Bonfrate L, Wang DQ and Portincasa P, *Liver steatosis, gut-liver axis, microbiome and environmental factors. A never-ending bidirectional cross-talk.* Journal of clinical medicine, 2020. **14;9**.
12. Tripathi A, D.J., Brenner DA, Karin M, Loomba R, Schnabl B and Knight R, *The gut–liver*

axis and the intersection with the microbiome. Nature reviews Gastroenterology & hepatology, 2018. **15**: p. 397-411.

13. Tobin NP, F.T., De Petris L and Bergh J, *The importance of molecular markers for diagnosis and selection of targeted treatments in patients with cancer.* Journal of internal medicine, 2015. **278**: p. 545-70.
14. Xu XR, H.J., Xu ZG, Qian BZ, Zhu ZD, Yan Q, Cai T, Zhang X, Xiao HS, Qu J and Liu F, *Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver.* Proceedings of the National Academy of Sciences, 2001. **98**.
15. Kunadirek P, P.N., Nookaew I, Tangkijvanich P and Chuaypen N, *Transcriptomic analyses reveal long non-coding RNA in peripheral blood mononuclear cells as a novel biomarker for diagnosis and prognosis of hepatocellular carcinoma.* International Journal of Molecular Sciences, 2022. **23**.
16. Ott JJ, S.G., Groeger J and Wiersma ST, *Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity.* Vaccine, 2012. **30**: p. 2212-2219.
17. Timothy R.Morgan, S.M.a.M.M.J., *Alcohol and hepatocellular carcinoma.* Gastroenterology, 2004. **127**: p. 87-96.
18. Westland C, D.I.W., Yang H, Chen SS, Marcellin P, Hadziyannis S, Gish R, Fry J, Brosgart C, Gibbs C and Miller M, *Hepatitis B virus genotypes and virologic response in 694 patients in phase III studies of adefovir dipivoxil.* Gastroenterology, 2003. **125**.
19. Odemuyiwa SO, M.M., Oyedele OI, Ola SO, Odaibo GN, Olaleye DO and Muller CP, *Phylogenetic analysis of new hepatitis B virus isolates from Nigeria supports endemicity of genotype E in West Africa.* Journal of medical virology, 2001. **65**.
20. Arauz‐Ruiz P, N.H., Visoná KA and Magnius LO, *Genotype F prevails in HBV infected patients of hispanic origin in Central America and may carry the precore stop mutant.* Journal of medical virology, 1997. **51**.
21. Sanchez LV, M.M., Bastidas‐Ramirez BE, Norder H and Panduro A, *Genotypes and S‐gene variability of Mexican hepatitis B virus strains.* Journal of medical virology, 2002. **68**: p. 24-32.

22. Kramvis A, K.M.a.F.G.H.B.v.g.V.-. *Hepatitis B virus genotypes.* Vaccine, 2005. **23**: p. 2409-23.

23. Louisirirotchanakul S, O.C., Arunkaewchaemsri P, Poovorawan Y, Kanoksinsombat C, Thongme C, Sa-nguanmoo P, Krasae S, Theamboonlert A, Oota S and Fongsatitkul L, *The distribution of hepatitis B virus genotypes in Thailand.* Journal of medical virology, 2012. **84**: p. 1541-7.

24. NN, Z., *Clinical significance of hepatitis C virus genotypes.* Clinical microbiology reviews, 2000. **13**: p. 223-35.

25. Messina JP, H.I., Flaxman A, Brown A, Cooke GS, Pybus OG and Barnes E, *Global distribution and prevalence of hepatitis C virus genotypes.* Hepatology, 2015. **61**: p. 77-87.

26. Wasitthankasem R, P.N., Treesun K, Posuwan N, Vichaiwattana P, Auphimai C, Thongpan I, Tongsima S, Vongpunsawad S and Poovorawan Y, *Prevalence of hepatitis C virus in an endemic area of Thailand: burden assessment toward HCV elimination.* The American Journal of Tropical Medicine and Hygiene, 2020. **103**: p. 175.

27. Bosch FX, R.J., Díaz M and Cléries R, *Primary liver cancer: worldwide incidence and trends.* Gastroenterology, 2004. **127**: p. 5-16.

28. Trépo, C., Henry LY Chan and Anna Lok, *Hepatitis B virus infection.* The Lancet, 2014: p. 2053-2063.

29. Tiollais, P., Christine Pourcel and Anne Dejean, *The hepatitis B virus.* Nature, 1985: p. 489-495.

30. Lamontagne, R.J., Sumedha Bagga and Michael J. Bouchard, *Hepatitis B virus molecular biology and pathogenesis.* Hepatoma research, 2016. **163**.

31. Baumert TF, J.F., Ono A and Hoshida Y, *Hepatitis C-related hepatocellular carcinoma in the era of new generation antivirals.* BMC medicine, 2017. **15**.

32. Petruzziello A, M.S., Loquercio G, Cozzolino A and Cacciapuoti C, *Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes.* World journal of gastroenterology, 2016. **22**: p. 7824.

33. Manns MP, B.M., Gane ED, Pawlotsky JM, Razavi H, Terrault N and Younossi Z, *Hepatitis C virus infection.* Nature reviews Disease primers, 2017. **3.**: p. 1-9.

34. Penin F, D.J., Rey FA, Moradpour D and Pawlotsky JM, *Structural biology of hepatitis C*

virus. Hepatology, 2004. **39**: p. 5-19.

35. Takamizawa A, M.C., Fuke I, Manabe S, Murakami S, Fujita J, Onishi E, Andoh T, Yoshida I and Okayama H, *Structure and organization of the hepatitis C virus genome isolated from human carriers.* Journal of virology, 1991. **65**: p. 1105-13.
36. P, S., *The origin of hepatitis C virus.* Hepatitis C virus: from molecular virology to antiviral therapy, 2013: p. 1-5.
37. WW, R.B.a.H., *Cellular and molecular interactions in coinfection with hepatitis C virus and human immunodeficiency virus.* Expert reviews in molecular medicine, 2008. **10**.
38. Nagaoki Y, H.H., Aikata H, Tanaka M, Naeshiro N, Nakahara T, Honda Y, Miyaki D, Kawaoka T, Takaki S and Hiramatsu A, *Recent trend of clinical features in patients with hepatocellular carcinoma.* Hepatology research, 2012. **42**: p. 368-75.
39. Y, N.H.a.O., *Non-B, non-C hepatocellular carcinoma.* International Journal of Oncology., 2013. **43**: p. 1333-42.
40. SA, T.D.a.H., *Nonalcoholic steatohepatitis and noncirrhotic hepatocellular carcinoma: fertile soil.* InSeminars in liver disease, 2012. **32**: p. 030-038.
41. Bugianesi E, L.N., Vanni E, Marchesini G, Brunello F, Carucci P, Musso A, De Paolis P, Capussotti L, Salizzoni M and Rizzetto M, *Expanding the natural history of nonalcoholic steatohepatitis: from cryptogenic cirrhosis to hepatocellular carcinoma.* Gastroenterology, 2002. **123**: p. 134-40.
42. Calle EE, R.C., Walker-Thurmond K and Thun MJ, *Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults.* New England Journal of Medicine, 2003. **348**: p. 1625-38.
43. Yasui K, H.E., Tokushige K, Koike K, Shima T, Kanbara Y, Saibara T, Uto H, Takami S, Kawanaka M and Komorizono Y, *Clinical and pathological progression of non-alcoholic steatohepatitis to hepatocellular carcinoma.* Hepatology research, 2012. **42**: p. 767-73.
44. AM, B.E.a.D.B., *Diagnosis of hepatocellular carcinoma.* Hpb, 2005. **7**: p. 26-34.
45. Choi JY, L.J.a.S.C., *CT and MR imaging diagnosis and staging of hepatocellular carcinoma: part II. Extracellular agents, hepatobiliary agents, and ancillary imaging features.* Radiology, 2014. **173**: p. 30.
46. Labgaa I, V.A., Dormond O, Demartines N, Melloul E, *The role of liquid biopsy in*

hepatocellular carcinoma prognostication. Cancers, 2021. **13**: p. 659.

47. von Felden J, G.-L.T., Schulze K, Losic B and Villanueva A, *Liquid biopsy in the clinical management of hepatocellular carcinoma.* Gut, 2020. **69**: p. 2025-34.

48. Jorge a. Marrero, L.M.K., Claude B. Sirlin, Andrew X. Zhu, Richard S. Finn, Michael M. abecassis, Lewis R. Roberts, and Julie K. Heimbach, *Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases.* Hepatology, 2018. **68**: p. 723-750.

49. José D. Debes, P.A.R., Jhon Prieto, Marco Arrese, Angelo Z. Mattos, André Boonstra and on behalf of the ESCALON Consortium, *Serum Biomarkers for the Prediction of Hepatocellular Carcinoma.* Cancers, 2021. **13**: p. 1681.

50. Zhiyi Han, W.F., Rui Hu, Qinyu Ge, Wenfeng Ma, Wei Zhang, Shaomin Xu, Bolin Zhan, Lai Zhang, Xinfeng Sun and Xiaozhou Zhou, *RNA seq profling reveals PBMC RNA as a potential biomarker for hepatocellular carcinoma.* Scientifc Reports, 2021. **11**: p. 17797.

51. Vahdat Poortahmasebi, A.N., Mohammad Foad Abazari, Mohsen Nasiri Toosi, Azam Ghaziasadi, Nader Mohammadzadeh, Ahmad Tavakoli, Azam Khamseh, Navid Momenifar, Omid Gholizadeh, Mehdi Norouzi and Seyed Mohammad Jazayeri, *Identifying Potential New Gene Expression-Based Biomarkers in the Peripheral Blood Mononuclear Cells of Hepatitis B-Related Hepatocellular Carcinoma.* Canadian Journal of Gastroenterology and Hepatology, 2022.

52. Pattapon Kunadirek , C.A., Supachaya Sriphoosanaphan, Nutcha Pinjaroen, Pongserath Sirichindakul, Intawat Nookaew, Natthaya Chuaypen & Pisit Tangkijvanich, *Identifcation of BHLHE40 expression in peripheral blood mononuclear cells as a novel biomarker for diagnosis and prognosis of hepatocellular carcinoma.* Scientifc Reports, 2021. **11**: p. 11201.

53. Jiang JW, C.X., Ren Z and Zheng SS, *Gut microbial dysbiosis associates hepatocellular carcinoma via the gut-liver axis.* Hepatobiliary & Pancreatic Diseases International, 2019. **18**: p. 19-27.

54. N, O.N.a.K., *Role of the gut–liver axis in liver inflammation, fibrosis, and cancer: a special focus on the gut microbiota relationship.* Hepatology Communications, 2019. **3**: p. 456-70.

55. Zhigang Ren, A.L., Jianwen Jiang, Lin Zhou, Zujiang Yu, Haifeng Lu, Haiyang Xie,

Xiaolong Chen, Li Shao, Ruiqing Zhang, Shaoyan Xu, Hua Zhang, Guangying Cui, Xinhua Chen, Ranran Sun, Hao Wen, Jan P Lerut, Quancheng Kan, Lanjuan Li, and Shusen Zheng, *Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma.* Gut microbiota, 2019. **68**: p. 1014-1023.

56. Sharpton SR, S.B., Knight R and Loomba R, *Current concepts, opportunities, and challenges of gut microbiome-based personalized medicine in nonalcoholic fatty liver disease.* Cell metabolism, 2021. **33**: p. 21-32.

57. HJ, L.P.a.F., *Development of a semiquantitative degenerate Real-Time PCR-based assay for estimation of numbers of butyryl-coenzyme A (CoA) CoA transferase genes in complex bacterial samples.* Applied and environmental microbiology, 2007. **73**: p. 2009-12.

58. Ren Z, J.J., Xie H, Li A, Lu H, Xu S, Zhou L, Zhang H, Cui G, Chen X and Liu Y, *Gut microbial profile analysis by MiSeq sequencing of pancreatic carcinoma patients in China.* Oncotarget, 2017. **8**: p. 95176.

59. Ponziani FR, B.S., Castelli C, Putignani L, Rivoltini L, Del Chierico F, Sanguinetti M, Morelli D, Paroni Sterbini F, Petito V and Reddel S, *Hepatocellular carcinoma is associated with gut microbiota profile and inflammation in nonalcoholic fatty liver disease.* Hepatology, 2019. **69**: p. 107-20.

60. Chen Z, H.P., Hui M, Yeoh YK, Wong PY, Chan MC, Wong MC, Ng SC, Chan FK and Chan PK, *Impact of preservation method and 16S rRNA hypervariable region on gut microbiota profiling.* Msystems, 2019. **4**.

61. W, N.K.a.K., *Big data and machine learning algorithms for health-care delivery.* The Lancet Oncology, 2019. **20**: p. 262-73.

62. Cai Z, P.R., Liu J and Zhong Q., *Machine learning for multi-omics data integration in cancer.* Iscience, 2022. **22**.

63. Hechen Huang, Z.R., Xingxing Gao, Xiaoyi Hu, Yuan Zhou, Jianwen Jiang, Haifeng Lu, Shengyong Yin, Junfang Ji, Lin Zhou and Shusen Zheng, *Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma.* Genome Medicine, 2020. **102**.

64. Somboon K, S.S.a.V.R., *Epidemiology and survival of hepatocellular carcinoma in the central region of Thailand.* Asian Pacific Journal of Cancer Prevention, 2014. **15**: p. 3567-

70.

65. MS, D.N.a.K., *Sample size calculator for comparing two paired means.* Statulator: An online statistical calculator, 2014.

66. Dore J, E.S.D., Levenez F, Pelletier E, Alberti A, Bertrand L, Bork P, Costea P.I, Sunagawa S, Guarner F,, S.A. Manichanh C, Zhao L, Shen J, Zhang C, Versalovic J, Luna R.A, Petrosino J, Yang H, Li S, Wang J,, and G.G. Allen-Vercoe E, Singh B. and IHMS Consortium, *IHMS-SOP 03V2: standard operating procedure for fecal sample self-collection, laboratory analysis handled within 4 to 24 hours (4 hours< x <24 hours).* International Human Microbiome Standards, 2015.

67. Klindworth A, P.E., Schweer T, Peplies J, Quast C, Horn M and Glöckner FO, *Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.* Nucleic acids research, 2013. **41**.

68. Ewels, P.A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen, *The nf-core framework for community-curated bioinformatics pipelines.* Nature biotechnology, 2020: p. 276-278.

69. Daniel Straub, A.P., Daniel Lundin, Jeanette Tångrot, emnilsson, DiegoBrambilla, nf-core bot, Asaf Peer, Gisela Gabernet, Venkat Malladi, PhilPalmer, Maxime U. Garcia, Phil Ewels, Colin Davenport, Harshil Patel and Kevin Menden, *nf-core/ampliseq: Ampliseq Version 2.3.0 (2.3.0).* Zenodo, 2022.

70. Quast C, P.E., Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J and Glöckner FO, *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic acids research, 2012. **41**: p. 590-6.

71. Douglas GM, M.V., Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C and Langille MG, *PICRUSt2 for prediction of metagenome functions.* Nature biotechnology, 2020. **38**: p. 685-8.

72. Pertea M, K.D., Pertea GM, Leek JT and Salzberg SL, *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nature protocols, 2016. **11**: p. 1650-67.

73. Shifu Chen, Y.Z., Yaru Chen and Jia Gu, *fastp: an ultra-fast all-in-one FASTQ*

preprocessor. Bioinformatics, 2018: p. 884-890.

74. Kim D, P.J., Park C, Bennett C and Salzberg SL, *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature biotechnology, 2018. **37**: p. 907-15.

75. Pertea M, P.G., Antonescu CM, Chang TC, Mendell JT and Salzberg SL, *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature biotechnology, 2015. **33**: p. 290-5.

76. Zhou Y, Z.B., Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C and Chanda SK, *Metascape provides a biologist-oriented resource for the analysis of systems-level datasets.* Nature communications, 2019. **10**.

77. Sherman BT, H.M., Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T and Chang W, *DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update).* Nucleic acids research, 2022. **23**.

78. Pedregosa F, V.G., Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V and Vanderplas J, *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-30.

79. Zhang Q, H.Y., Luo N, Patel SJ, Han Y, Gao R, Modak M, Carotta S, Haslinger C, Kind D and Peet GW, *Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma.* Cell, 2019. **179(4)**: p. 829-845.

80. Liu Q, L.F., Zhuang Y, Xu J, Wang J, Mao X, Zhang Y and Liu X, *Alteration in gut microbiota associated with hepatitis B and non-hepatitis virus related hepatocellular carcinoma.* Gut pathogens, 2019. **11(1)**: p. 1-3.

81. Fianchi F, L.A., Gasbarrini A, Grieco A and Miele L, *Nonalcoholic Fatty Liver Disease (NAFLD) as Model of Gut–Liver Axis Interaction: From Pathophysiology to Potential Target of Treatment for Personalized Therapy.* International Journal of Molecular Sciences, 2020. **22(12)**: p. 6485.

82. T, M.D.a.P., *Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism.* Gut microbes, 2016. **7(3)**: p. 189-200.

83. Nastasi C, C.M., Bonefeld CM, Geisler C, Hansen M, Krejsgaard T, Biagi E, Andersen MH, Brigidi P, Ødum N and Litman T, *The effect of short-chain fatty acids on human monocyte-derived dendritic cells.* Scientific reports, 2015. **5(1)**: p. 1-10.

84. Li P, Z.Z., Gong J, Zhang Y and Zhu X, *S-Adenosylmethionine attenuates lipopolysaccharide-induced liver injury by downregulating the Toll-like receptor 4 signal in Kupffer cells.* Hepatology international, 2014. **8**: p. 275-284.

85. Guo J, C.L., Luo N, Li C, Chen R, Qu X, Liu M, Kang L and Cheng Z, *LPS/TLR4-mediated stromal cells acquire an invasive phenotype and are implicated in the pathogenesis of adenomyosis.* Scientific reports, 2016. **6(1)**: p. 1-10.

86. Killeen SD, W.J., Andrews EJ and Redmond HP, *Bacterial endotoxin enhances colorectal cancer cell adhesion and invasion through TLR-4 and NF-ΚB-dependent activation of the urokinase plasminogen activator system.* British Journal of Cancer, 2009. **100(10)**: p. 1589-1602.

87. Komiyama S, Y.T., Takemura N, Kokudo N, Hase K and Kawamura YI, *Profiling of tumour-associated microbiota in human hepatocellular carcinoma.* Scientific reports, 2021. **11(1)**: p. 1-9.

88. Henke MT, K.D., Cassilly CD, Vlamakis H, Xavier RJ and Clardy J, *Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide.* Proceedings of the National Academy of Sciences, 2019. **116(26)**: p. 12672-12677.

89. Zuhl M, S.S., Lanphere K, Conn C, Dokladny K and Moseley P, *Exercise regulation of intestinal tight junction proteins.* British journal of sports medicine, 2014. **4(12)**: p. 980-986.

90. Y, V.C.a.L., *A Comparative Review of Toll-Like Receptor 4 Expression and Functionality in Different Animal Species.* Frontiers in Immunology, 2014. **5**: p. 96623.

91. Płóciennikowska A, H.-J.A., Borzęcka K and Kwiatkowska K, *Co-operation of TLR4 and raft proteins in LPS-induced pro-inflammatory signaling.* Cellular and molecular life sciences, 2015. **72**: p. 557-581.

92. Haynie T, G.S., Drees C, Heaton T, Mitton T, Gleave Q, Bendelac A, Deng S and Savage PB, *Synthesis of the pentasaccharide repeating unit from Ruminococcus gnavus and measurement of its inflammatory properties.* RSC advances, 2021. **11(24)**: p. 14357-61.

VITA

NAME	Jakkrit Khamjerm
DATE OF BIRTH	11 August 1995
PLACE OF BIRTH	Nakhonsawan
INSTITUTIONS ATTENDED	Srinakharinwirot University
HOME ADDRESS	294/672 Ideo Charan 70 Riverview, Thanon Charan Sanit Wong, Khwaeng Bang Phlat, Khet Bang Phlat, Krung Thep Maha Nakhon 10700, Thailand