CLASSIFICATION OF ORGANOID TRANSCRIPTOMIC PROFILES UNRAVELLING COLORECTAL CANCER MOLECULAR SUBTYPES

Miss Pattarin Nuwongsri

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Medical Sciences
Common Course
FACULTY OF MEDICINE
Chulalongkorn University
Academic Year 2020

การศึกษารูปแบบการแสดงออกของยีนในกลุ่มเซลล์สามมิติเพื่อใช้ในการจัดกลุ่มทางโมเลกุลของมะเร็งลำไส้ใหญ่

น.ส.ภัทริน นุวงศ์ศรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์การแพทย์ ไม่สังกัดภาควิชา/เทียบเท่า
คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	CLASSIFICATION OF ORGANOID TRANSCRIPTOMIC PROFILES UNRAVELLING COLORECTAL CANCER MOLECULAR SUBTYPES
By	Miss Pattarin Nuwongsri
Field of Study	Medical Sciences
Thesis Advisor	Assistant Professor NIPAN ISRASENA, Ph.D.
Thesis Co Advisor	SIRA SRISWASDI, Ph.D.

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

..... Dean of the FACULTY OF MEDICINE
(Professor SUTTIPONG WACHARASINDHU, M.D.)

THESIS COMMITTEE

..... Chairman
(Professor VILAI CHENTANEZ, Ph.D.)

..... Thesis Advisor
(Assistant Professor NIPAN ISRASENA, Ph.D.)

..... Thesis Co-Advisor
(SIRA SRISWASDI, Ph.D.)

..... Examiner
(Professor Duangporn Werawatganon, M.D.)

..... External Examiner
(Sawannee Sutheeworapong, D.sc.)

ภัทริน นุวงศ์ศรี : การศึกษารูปแบบการแสดงออกของยีนในกลุ่มเซลล์สามมิติเพื่อใช้ในการจัดกลุ่มทางโมเลกุลของมะเร็งลำไส้ใหญ่. (CLASSIFICATION OF ORGANOID TRANSCRIPTOMIC PROFILES UNRAVELLING COLORECTAL CANCER MOLECULAR SUBTYPES) อ.ที่ปรึกษาหลัก : รศ.ดร. นพ.นิพัญจน์ อิศรเสนา ณ อยุธยา, อ.ที่ปรึกษาร่วม : ดร.สิระ ศรีสวัสดิ์

มะเร็งลำไส้ใหญ่เป็นโรคที่มีความหลากหลายทางพันธุกรรม ทำให้มีการแสดงออกของยีนที่แตกต่างกันไป ก่อนหน้านี้มีงานวิจัยที่ทำการจัดกลุ่มทางโมเลกุลของมะเร็งชนิดนี้ด้วย Consensus Molecular Subtype (CMS) ซึ่งมีแนวโน้มในการนำไปใช้ทำนายการดำเนินโรคและการตอบสนองต่อการรักษา แต่อย่างไรก็ดี CMS ใช้ข้อมูลการแสดงออกของยีนจากชิ้นเนื้อที่ประกอบด้วยเซลล์มะเร็งและเซลล์อื่นๆภายใน stromal เช่น เซลล์เม็ดเลือดขาว จึงนำมาสู่คำถามวิจัยทีว่าหากใช้ข้อมูล Transcriptome ของเซลล์มะเร็งโดยเฉพาะจะช่วยให้สามารถแยกกลุ่มของมะเร็งลำไส้ใหญ่ได้ดีขึ้นหรือไม่ และนำมาสู่การศึกษานี้ที่จำแนกกลุ่มของมะเร็งลำไส้ใหญ่ โดยใช้ข้อมูลการแสดงออกของยีนที่ได้จากกลุ่มเซลล์สามมิติ (Organoid) ซึ่งประกอบขึ้นจากเซลล์มะเร็งเป็นหลัก ในการศึกษานี้เริ่มต้นจากการประเมินความเปลี่ยนแปลงในระดับ Genome โดยใช้ข้อมูล whole exome sequencing จากผลการวิเคราะห์พบลักษณะของ chromosomal instability (CIN) และ microsatellite instability (MSI) ซึ่งเป็นกระบวนการในการเกิดมะเร็งลำไส้ใหญ่ที่แตกต่างกัน นอกจากนั้นข้อมูลนี้ยังแสดงให้เห็นลักษณะการกลายพันธุ์ที่หลากหลายของยีนที่มีความเกี่ยวข้องกับมะเร็งชนิดนี้ ต่อมาในส่วนของข้อมูล Transcriptome ของ organoid นั้นสามารถจำแนกได้เป็น 4 กลุ่ม (P1-P4) โดยใช้วิธีการ non-negative matrix factorization (NMF) ซึ่งยีนที่เป็นเอกลักษณ์ของแต่ละกลุ่ม แสดงให้เห็นว่าแต่ละกลุ่มนั้นมีลักษณะเฉพาะที่แตกต่างกัน P1มีกระบวนการเมทาบอลิซึมของไขมันและคอเลสเตอรอลที่สูงขึ้น P2และP3 มีการแสดงออกของ TGFβ pathway ที่สูง และกลุ่มสุดท้าย P4 ที่แสดงคุณสมบัติคล้ายกับเซลล์ต้นกำเนิดและมีการแสดงออกของยีนในกลุ่มที่มีหน้าที่เกี่ยวกับการซ่อมแซม DNA ที่สูงขึ้น ซึ่งลักษณะดังกล่าวนี้มีความสอดคล้องกับการต้านทานต่อยาเคมีบำบัดและรังสีรักษา นอกจากนั้น ribosome biogenesis pathway ยังถูกกระตุ้นมากขึ้นในกลุ่ม P4 ซึ่งลักษณะดังกล่าวอาจจะสามารถนำมาพัฒนาเป็นเป้าหมายในการรักษามะเร็งลำไส้ใหญ่ได้ในอนาคต ต่อมาเพื่อหายีนที่เป็นตัวบ่งชี้ของแต่ละกลุ่ม LASSO logistic regression model จึงถูกสร้างขึ้นเพื่อหายีนที่สามารถจำแนกแต่ละกลุ่มได้ จากผลการศึกษานี้แสดงให้เห็นว่ายีนเอกลักษณ์ของกลุ่มเซลล์สามมิตินั้นจึงอาจจะนำมาพัฒนาให้เป็นเครื่องมือสำหรับแยกกลุ่มและพัฒนาการรักษาสำหรับมะเร็งลำไส้ใหญ่ที่มีความจำเพาะมากขึ้น

สาขาวิชา	วิทยาศาสตร์การแพทย์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2563	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6174019030 : MAJOR MEDICAL SCIENCES

KEYWORD: colorectal cancer, molecular classification, organoid, non-negative matrix factorization (NMF), gene signature

Pattarin Nuwongsri : CLASSIFICATION OF ORGANOID TRANSCRIPTOMIC PROFILES UNRAVELLING COLORECTAL CANCER MOLECULAR SUBTYPES. Advisor: Asst. Prof. NIPAN ISRASENA, Ph.D. Co-advisor: SIRA SRISWASDI, Ph.D.

Colorectal cancer (CRC) is genetically and transcriptomically heterogeneous disease. Molecular subtyping of colorectal cancer using consensus molecular subtype (CMS) system demonstrated the potential predictive value for tumor progression and treatment response. However, the CMS system was developed from data of whole tissues containing both cancer and non-tumor transcripts components for classification which does not directly represent intrinsic heterogeneity of cancer cells. In this study genetic profiles of CRC organoids were investigated first, and the results indicate chromosomal instability (CIN) and microsatellite instability (MSI) as pathogenic pathways of CRC. Furthermore, the results also revealed diverse patterns of somatic mutations of these CRC organoids. Subsequently, we evaluated a strategy of subtyping CRCs based on transcriptomics data from patient-derived CRC organoids, which mainly contain cancer cells. We demonstrated that using non-negative matrix factorization (NMF) CRC cancer organoids could be classified into four groups (P1-P4). Cluster-specific genes and Gene Set Enrichment Analysis (GSEA) displayed different characteristics of each group. P1 exhibit enriched lipid and cholesterol metabolism pathways and P2 and P3 presented high TGF-β pathway. Lastly, P4 show stem cell-like properties and highly expressed genes in the DNA repair pathway associated with chemotherapy and radiation resistance. Moreover, P4 organoids present a hyperactivated ribosome biogenesis pathway which may be developed as a biomarker of P4 and a target of CRC treatment. Then, LASSO logistic regression was built to identify gene signatures and developed a classifier of each group of organoids. These results suggested that the signature gene of organoid groups has the potential to be developed into a useful tool for CRC subtyping and developing more specific therapeutic strategies.

Field of Study:	Medical Sciences	Student's Signature .....
Academic Year:	2020	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support and nurturing of many people. First, I would like to express the deepest appreciation to my advisor. Associate Professor Nipan Isarasena Na-ayuthaya, M.D. Ph.D. for valuable suggestions and supportive guidance. Then, I would like to extend my sincere thanks to my Co-advisor, Dr. Sira Srisawasdi, for inspiring me to be interested in the field of computational molecular biology. Likewise, his expertise was invaluable in formulating the research questions and methodology. In addition, his insightful feedback pushed me to sharpen my thinking. Moreover, I would like to express my deepest appreciation to Dr. Praewphan Inrunruanglert not only for her assistance with the collection of my data but also her advice and support. Besides, I would like to express my deepest appreciation to my committee including Prof. Vilai Chentanez, Ph.D., Prof. Duangporn Werawataganon, M.D., and Dr. Sawannee Sutheeworapong, D.sc. for their comments and suggestions.

My special thanks are extended to my friends in the Stem cells therapy research unit and the computational molecular biology (CMB) group, King Chulalongkorn Memorial hospital and Chulalongkorn university that being supporting me in the way they could.

Finally, I would like to express my gratitude to my family for their support and encouragement throughout my study.

Pattarin Nuwongsri

TABLE OF CONTENTS

Table

Page

Figure

Page

LIST OF ABBREVIATIONS

Abbreviation	Definitions
CIN	Chromosomal instability
CMS	Consensus Molecular Subtype
CNV	Copy-number variation
CRC	Colorectal cancer
CRIS	Colorectal Cancer Intrinsic Subtype
DDR	DNA damaged repair
DEG	differentially expressed gene
FA	Fanconi anemia
FDR	False Discovery Rate
GATK	Genome Analysis Toolkits
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MMR	Mismatch repair
MSI	Microsatellite instability
NMF	Non-negative Matrix Factorization
NOR	Nucleolar Organizer Region
PBMC	Peripheral blood mononuclear cell
PDO	Patient-derived organoid
TCGA	The Cancer Genome Atlas
WES	Whole Exome Sequencing

CHAPTER I

INTRODUCTION

1.1 Background and Rationale

Colorectal cancer (CRC) is the third most common cancer worldwide and frequently diagnosed at advanced clinical stage (1). 60-65% of CRC sporadically arise through acquired somatic mutations and epigenetic alterations. Importantly, the 5-year overall survival rate drops drastically from 64% to 14% when the tumor becomes metastatic (2). Although an increasing number of therapeutic treatments have been developed for CRC, clinical outcomes are still undesirable. This is because the disease is highly heterogeneous and can progress through many alternative pathways, each with different genetic alterations, molecular profiles, clinical outcomes, and treatment responses. Consequently, it is challenging to identify the optimal therapy for each patient. Current clinical classifications of CRC depend on histopathological features and a simplistic tumor-node-metastasis (TNM) staging. However, patients with the same stage respond vastly differently to the same treatment. For these reasons, improvements to CRC subtype classification and treatment response prediction are needed (3). While additional molecular markers, such as microsatellite instability (MSI) status and BRAF and KRAS mutations, have been introduced, they could not capture the complexity of CRC tumor biology and are insufficient for treatment selection or prognosis prediction (3, 4).

Gene expression profiling technologies such as microarray and RNA-sequencing can provide comprehensive molecular characteristics of a tumor. Accordingly, CRC classification framework has recently shifted towards transcriptomic data (4-8). A major breakthrough came in 2015 when a network-based analysis was used to unify six CRC classification studies and derive the first consensus CRC subtyping scheme, named the consensus molecular subtype (CMS). CMS method stratifies patients based on gene expression data into four major subtypes (CMS1-4) and a separate group of patients with mixed phenotypes. It was also speculated that further refinements of CRC classification, such as segregation of intra-CMS subgroups and delineation of unclassified samples with mixed phenotypes, will be necessary. One possible area of improvement is to reduce the interference from non-tumor cells, such as stromal components and infiltrated immune cells, in the gene expression profile of bulk tumor tissue. To avoid effects of stromal components, a CRC intrinsic subtypes (CRIS) classification scheme has been developed

by implanting patient-derived xenografts (PDXs) into mice and subsequently extracting human-specific transcriptomic profiles from the PXDs. This technique enables acquisition of tumor cell-specific gene expression data. An unsupervised clustering of PXD-based transcriptomic data indicated that CRC tumor may be stratified into five molecular classes (CRIS-A to CRIS-E) (9). As expected, a finer classification of CRC subtypes, especially for CMS2 group, can be achieved when interference from non-tumor cells was reduced. However, PDX-based classification method still suffers from cross-species reactivity between human and mouse cytokines which distort cancer cell transcriptome and from some stromal-derived transcripts. Inconsistencies between CMS and CRIS approaches also need to be explained. Therefore, alternative methods for extracting cancer cell's transcriptional profile from patient tumors are needed to delicately stratify CRC subgroups (9).

Recently, 3D cell culture systems have been developed. These techniques allow us to grow organoids composing of multiple organ-specific epithelium in the absence of stromal cells. Furthermore, organoids can preserve intra-tumoral heterogeneity, transcriptomic pattern, and key phenotypes of the original tissue (10, 11). Patient-derived organoids (PDOs) serve as effective preclinical models of human cancer as well as enable rapid, high-throughput ex vivo drug testing and screening since PDOs could be propagated and expanded within a few weeks. Notably, drug response of PDOs have been shown to correlate with the patients' actual response (10). Hence, we hypothesize that intrinsic transcriptomic profile of cancer cells could be gleaned through gene expression data of CRC organoids which consist mainly of epithelial cells.

In this study, transcriptomic data from 54 PDOs of CRC patients were grouped into 4 prospective subtypes by unsupervised clustering methods. Then, differential expression and functional enrichment analyses were performed to identify molecular signatures of each group. Furthermore, copy-number variation, MSI score and mutation profiles of CRC organoids were identified using exome sequencing. Interestingly, organoids with radiation resistance were clustered together and demonstrated shared chromosomal instability and upregulation of DNA repair and ribosome biogenesis pathways. Finally, supervised machine learning techniques were used to construct a subtype classification model and to identify signature genes that contribute to the classification of each subtype.

1.2 Research question

Whether transcriptomic data of organoids could be used to classify molecular subtype of CRC?

1.3 Hypothesis

The gene expression profiles of colorectal cancer organoids provide molecular subtype of CRC and gene signature of each organoid group

1.4 Objectives

- To classify molecular subtype of CRC by using transcriptomic data of organoids
- To identify gene markers specific to each organoid group

CHAPTER II

LITERATURE REVIEWS

2.1 Stepwise progressions of colorectal cancer

Tumorigenesis of colorectal cancer progresses through three different pathways including adenoma-carcinoma sequence, Serrated pathway, and inflammatory pathway. Adenoma-carcinoma sequence is a classical or canonical pathway of CRC. This pathway begins with the acquisition of adenomatous polyposis coli (APC) mutations that upregulate Wnt/β-catenin signaling pathway, followed by KRAS mutation activation and TP53 tumor suppressor gene inactivation. Furthermore, transformation into metastatic phenotypes also occurs through dysregulation of multiple signaling pathways involved in cell cycle regulation and cellular proliferation. Chromosomal instability (CIN) due to loss of heterozygosity (LOH) and aneuploidy have also been found in 85% of sporadic tumors.

Serrated pathway drives the progression from normal cells to hyperplastic polyp. It has been reported that serrated CRC patients has worse prognosis than patients with aberrations in canonical pathway. There are two characteristic molecular events in the serrated pathway. A critical early event is BRAF mutation which causes uncontrolled cell proliferation via activation of MAPK pathway and leads to hyperplastic polyp formation. Another event, called CpG island methylator phenotype (CIMP), is the hypermethylation of specific target promoter which contributes to microsatellite instability (MSI) and inactivation of tumor suppressor genes that promote later progression of polyps into sessile serrated adenoma and carcinoma. Notably, CIMP positivity was found about 75% of sessile serrated adenomas. Additionally, MSI is marked by alterations in the length of microsatellite (short nucleotides repeated and distributed along DNA sequence), this is owing to loss of DNA mismatch repair (MMR) system leading to genetic instability.

Chronic inflammation can also lead to carcinogenic progression. This pathway begins with no dysplasia unlike canonical adenoma and serrated adenoma. Instead, dysplasia subsequently arises on the background of chronic inflammation. This type of CRC is frequently located in flat mucosa which conceals the lesion. Major molecular events in this pathway consist of TP53 mutation in the early stage and rare APC mutations in the late stage. Less than 2% of all CRCs arise through this pathway (1).

2.2 Colorectal treatment

Generally, the ideal treatment of CRC is to entirely remove all tumors and metastases through surgery. However, this is not possible especially for advanced stage CRCs. Accordingly, radiotherapy and chemotherapy are used to halt the growth and spread of tumors in such patients (12). The standard chemotherapies for metastatic CRC utilize fluoropyrimidines, oxaliplatin and irinotecan, which result in median overall survival of approximately 18 to 20 months. Drugs such as epidermal growth factor receptor (EGFR) inhibitors can also be prescribed together with chemotherapies to improve the median survival to 30 months. Several agents have been developed to target known CRC tumorigenesis and metastasis pathways, including Wnt/β-catenin, Notch, Hedgehog and TGF-β/SMAD. Some agents also target signaling cascades such as PI3K/AKT or RAS/RAF. At present, there is no proven CRC treatment that is effective for every patient.

2.3 Clinical classifications of colorectal cancer

The union for international cancer control (UICC) and American Joint Committee on Cancer (AJCC) suggested the widely used Tumor Node Metastasis (TNM) classification guidelines for determining colorectal cancer staging and selecting treatments. Yet, the treatment outcomes of CRC patients with the same TNM classifications are still highly variable (13).

To date, several mutation-based classifications have been used to guide treatment selections for CRCs. For example, TP53 mutations are predictive of decreased sensitivity to most chemotherapeutic agents, especially 5-fluorouracil. Previous studies found that BRAF inhibitors were ineffective in CRC patients with BRAF V600E mutations owing to EGFR feedback activation (4). MSI-high status is associated with poor response to 5-fluorouracil-based chemotherapy but suggests the possibility for immunotherapy with immune checkpoint targeting molecules such as PD-1. Although mutation-centered CRC classification has shown some promises in prognosis prediction and aiding treatment selection (1), it still does not provide sufficient predictive power and insight to improve our understanding of CRC tumor biology.

Table 1 the previous six independent studies used for the consensus molecular subtype identification (4)

Classification system	Discovery dataset	Validation dataset	Clustering method	Statistic for cluster count selection	Classification method	subtypes
Schlicker et al. **(2012) (14)**	62 samples	1643 samples	Iterative non-negative matrix factorization (NMF) -based consensus clustering	Cophenetic correlation coefficient	Two-step hierarchical clustering	Subtype 1.1, 1.2, 1.3 Subtype 2.1, 2.2
Marisa et al. **(2013) (15)**	443 samples	1029 samples	Classical consensus clustering	Area under cumulative distribution function (CDF) curve	Standard centroid-based classifier	C1 -C6
Sadanandam **et al. (2013) (16)**	445 samples	744 samples	NMF-based consensus clustering	Cophenetic correlation coefficient	Prediction analysis for microarrays (PAM)	Goblet-like, enterocyte, stem-like, inflammatory, transit-amplifying
De Sousa E Melo **et al. (2013) (17)**	90 samples	1074 samples	Classical consensus clustering	Gap statistic	Prediction analysis for microarrays (PAM)	CCS1-CCS3
Budinska **et al. (2013) (18)**	1113 samples	720 samples	Classical consensus clustering	Dynamic cut tree	Multiclass linear discriminant (LDA)	Surface crypt-like, lower crypt-like, CIMP-H-like, mesenchymal, mixed
Roepman **et al.(2014) (19)**	188 samples	543 samples	Hierarchical clustering	N/A	Single sample centroid-based classifier	Type A-C

2.4 Consensus molecular subtype (CMS) classification

Multiple molecular subtyping techniques were evaluated and resulted in inconsistent results thus to manage with it the CRC subtyping consortium (CRCSC) was formed. Subsequently, consensus molecular subtype (CMS) classification has been developed by using network-based approach on large-scale data from six independent studies of transcriptomic-based subtyping methods (18 CRC data sets, n = 4,151 patients) in order to study the association among these six classifications. CMS classification is able to classify most CRC tumors into four molecular subtypes with unique pathway enrichment traits.

Firstly, most MSI-high tumors are in the CMS1 (immune subtype, 14%) most tumor present hypermutation, hypermethylation and contain BRAF(V600E) mutations. Moreover, its present immune cell infiltration within tumor microenvironment which is significantly associated with better prognosis in MSI tumors. It is reported that local-infiltration is highly enriched with tumor-infiltrating cytotoxic T lymphocytes (CTLs) in core tumor area and surrounding peritumoral area. The local inflammatory response is widely reported in tumor progression in most of the cancers and presence of tumor infiltrating lymphocytes (TILs) are most important in the suppression of tumor progression and invasion. Inhibitors, such as immune checkpoint inhibitor that stimulate TILs have been proposed to regulate CRC progression such as PD1 blocker. Secondly, tumor with chromosomal instability (CIN), that are commonly non-hypermutated, can be transcriptome-based subclassified into three groups: CMS2 (canonical subtype, 37%); CMS3 (metabolic subtype, 13%); CMS4 (mesenchymal subtype, 23%). CMS2 tumors showed more frequent copy-number alterations than other subtypes. Additionally, found that WNT and MYC downstream targets are highly upregulated and higher expression of the EGFR, ERBB2 (also known as HER2), insulin-like growth factor 2 (IGF2), as well as cyclins. Moreover, CMS3 tumors are characterized by up to 30% present with MSI and gene hypermethylation in intermediate levels. It also contains metabolic reprogramming as well as it enriched for KRAS-activating mutations linked to marked metabolic adaptation in CRC. Therefore, an understanding of glucose metabolic pathway in cancer may also be seen as novel therapeutic targets. Finally, CMS4 tumors are activated in pathways associated with epithelial-mesenchymal transition (EMT) and stemness such as TGFβ and show prominent expression of proteins in extracellular matrix remodeling and angiogenesis. This subtype tends to be diagnosed at more advance stages. Corresponding to patient cohort CMS4 tumors result in worse overall survival. Importantly, the combination of chemotherapy and TGF-β receptor (TGFR) inhibitor has already moved to clinical trials in patients

whose tumor test positive for TGFβ activated. Instead, there are 13% of early-stage tumor cannot be assigned in any subtypes, demonstrating mixed phenotypes or intra-tumoral heterogeneity.

In the pre-clinical studies, they found an association with sensitivity to chemotherapy-induced apoptosis prevalent in CMS2 and CMS4 (20). Moreover, different studies retrospectively evaluated CMS as a prognostic factor for stage III CRC patients treated with FOLFOX adjuvant chemotherapy, finding that CMS was predictive in these patients. In 2019, Lenz et al. demonstrated that the CMSs are highly prognostic and predictive for overall survival (OS) and progression-free survival (PFS). In the CMS1 group, patients treated with bevacizumab had a significantly longer OS than those treated with cetuximab. For the CMS2, patients treated with cetuximab had a significantly longer OS than patients treated with bevacizumab. These findings highlight the possible application of CMSs in clinic and suggest that refinement of the CMS classification may provide a path toward identifying patients who are most likely to benefit from specific targeted therapy (21). Menter et al. purposed that absent knowledge of the CMS, multiple drugs have been tried on the entire CRC population and it may only show a low responses rate due to the drug affects to specific group of patients. Unfortunately, these drugs would likely have been discarded as ineffective for CRC. For these reasons, they assume that if we have a drug targeting pathway alteration which is a characteristic of each subtype. This can lead to increasing of response rate of these subgroups and greatly enhanced progression free and overall survival, this would be considered a complete success (22).

However, most of data used in training set were derived from bulk tumor tissue which provide transcriptomic profiles of stromal cell resulting in variation of expression patterns due to different location of tumor. Moreover, stromal transcripts are significantly influenced molecular classification processes. Besides, previous study suggested that it is necessary to perform further refinement in subtype classification with intra-CMS subgroup and better classification of samples with mixed phenotypes.

Recent study demonstrated that CMS2 subtype has the same proportion in both the early and advanced stages. This is possibly the most heterogeneous gene expression subtype. In fact, CMS2 includes two of their original CRCAssigner subtypes (enterocyte and transit-amplifying or TA) and three Marisa subtypes (C1, C5 and C6). Thus, it may be reasonable to subdivide the CMS2 to further understand biological heterogeneity, stage distribution, and potential personalized target of this subtype. Similarly, the recent study demonstrated significantly

different prognostic value when the CMS4 subtype was further subdivided into CMS4-C4 (worse DFS and OS) and CMS4-not C4 based on Marisa classification (8). These examples highlight how CMS subtypes define the overall profiles of major CRC subgroups; however, even within each subtype, there may be biological variability and important sub-subtypes with distinctive biological parameter that requires careful consideration(23). In 2019, Purcell et al. (24) investigated the utility of CMS to predict prognosis of CRC patients compared to the routinely used staging. They found that CMS4 was not an independent prognostic marker for survival while TNM staging significantly explains mortality independently of age and gender. Multiple studies revealed that intra-tumoral heterogeneity may affect the classification of CMS4 tumors due to the EMT-associated genes seen in CMS4 tumors may present upregulated gene derived from fibroblast and mesenchymal cells present in the stromal background rather than directly from the tumor itself. Moreover, previous studies suggested that the location and number of tumor biopsies can undermine the accuracy of CMS (5, 9, 25).

2.5 Colorectal cancer intrinsic subtype (CRIS) classification

It is necessary to classify patients by using transcriptomic data still unaffected by stromal variables. Thus, Colorectal cancer intrinsic subtype (CRIS) classification has been developed based on human-specific transcriptome in CRC PDX models because original tumor stroma is replaced by mouse stroma. Consequently, using human-specific probes can extract intrinsic gene expression of cancer cells. Then, transcriptomic patterns were analyzed through unsupervised clustering to stratify samples into five subgroups: (i) CRIS-A mostly are MSI tumor together with CpG island methylator phenotype (CIMP) and hypermutation as well as KRAS and BRAF mutation CRIS-A has mucinous and glycolytic phenotypes: (ii) CRIS-B contains BRAF mutations, displayed strong TGFβ activity and epithelial-mesenchymal transition (EMT) characteristics. They purposed that CRIS-B tumors had poor prognosis: (iii) CRIS-C shows KRAS-wild-type as well as contains MYC proto-oncogene and elevated EGFR signaling: (iv) CRIS-D was enriched for IGF2 amplification and WNT activation: (v) CRIS-E contains KRAS and TP53 mutations, Paneth cell-like phenotype. Additionally, CRIS-C, D and E are presented CIN. Importantly, molecular subtypes of previous studies have not reported to associated with these characteristics of individual CRIS subtypes. This indicates removing of stromal transcriptome throughout the classification process improved sensitivity to identification of intrinsic characteristics of cancer cells. Another study compared CMS to CRIS using multiple sampling method approaches, they concluded that CRIS provide more spatially, and temporally robust classification of molecular subtypes compared to CMS (26).

Thus, this group combined CRIS transcriptional subtyping and CD8 immunohistochemistry to identify poor prognosis stage II/III CRC patients who were able to benefit from adjuvant chemotherapy (27).

Previous study revealed that CRIS signature genes are predominantly expressed in epithelial cell type contribute to improve subgroup segregation and this method perform higher level of agreement in subtype classification than the CMS classifier, when perform the same data. However, they suggested that some stromal-derived expression patterns are remained in CRIS classification. Additionally, PDX models might present cross-species reactivity between human and mouse cytokines leading to distortion of cancer cell transcriptome. These data indicated that alternative methods to keep exclusively cancer cell transcriptional profile from patient tumors are necessary to delicately segregated subgroup (9).

2.6 Unsupervised clustering methods

According to the advent of microarray and RNA-sequencing, it is possible to simultaneous observe gene expression data of the sample. However, interpretation of the expression data to gain insight of biological process and disease mechanisms are still challenged. Thus, various methods have been developed for clustering genes or samples.

Hierarchical clustering (HC) has been developed for clustering genes or samples that show similar expression patterns. HC is a frequently used and beneficial method. It has been successfully used to analyze gene expression patterns to predict patient outcome among lymphoma patients (28) and to provide molecular portraits of breast cancer (29). However, this method has limitations in their ability to focus on the prevailing structures in a data set and fail to capture alternative structures and local behavior. Moreover, HC has the additional drawback that it imposes a stringent tree structure on the data, is highly sensitive to the metric used to assess similarity, and normally requires subjective evaluation to define clusters.

Non-negative matrix factorization (NMF) algorithm has been firstly proposed by Lee and Seung (30) as part-based learning of faces and semantic features of text. For example, NMF decompose human face images into parts reminiscent of features such as eye, nose, etc. This method is different to other methods such as principal components analysis that learn holistic representations. The NMF is discriminated from the other methods by its application of non-negative constraints. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. Next, several variations of it have been proposed

for clustering a single high-dimensional data. In 2004, Brunet et al. demonstrated the use of NMF to reduce the dimension of gene expression data to a small number of metagenes (31). Then, the metagene expression patterns provide a robust clustering of samples. This study suggested that the ability of NMF to retrieve meaningful biological information from microarray data of cancer. Notably, this method exhibit benefits over other methods such as hierarchical clustering. Moreover, it seems to less sensitive to gene selection or initial conditions and allow to detect different or context-dependent patterns of gene expression in complex biological systems. Thus, they proposed NMF as a general method for robust molecular pattern discovery.

2.7 Colorectal cancer organoid: a pre-clinical cancer model

In the past few decades traditional cancer cell lines and animal models have been used to study about tumorigenesis, tumor progression and drug responses of colorectal cancer. Whereas this approach is associated with a high failure rate of drug responses in the later clinical trial steps due to cancer cell lines poorly represent many features of the original tumors include genetic heterogeneity in the cancer cells lead to gene expression adaptations which attributed to varying treatment responses. Subsequently, Patient-derived xenografts (PDXs) model has been developed by isolating tumor cells from patients and transplanting them into immunodeficiency mice. This approach almost completely represents the genotypes and phenotypes of tumors. Nonetheless, PDX models are limited by a long time of establishment including inappropriate for purposes of high-throughput screening.

Organoids, 3D culture models, have been proposed as a pre-clinical cancer model based on knowledge of signal regulation of self-renewal, proliferation, and differentiation within intestinal stem cells (ISCs), allowing continuing expansion of Lgr5+ ISCs into crypt-villus structure. For CRC organoid propagation CRC biopsies were isolated and embedded within Matrigel along with combination of specific niche factors to mimic microenvironment in the crypt include epidermal growth factor (EGF), Noggin, R-spondin1 and Wnt3A with addition of TGFβ inhibitor and p38 inhibitor (11).

The CRC organoids consist of multiple cell types of the organ which can recapitulate the heterogeneity of their original tumors (32). Additionally, it contains the gene expression patterns and some key features and functions of that organs. Furthermore, many studies revealed that organoids demonstrated concordance of somatic mutations patterns between organoids and

corresponding biopsy (33, 34). Due to organoids can be generated and expanded from every individual patient and it can closely resemble to the original tumors indicate that organoid is a promising model which can more representative and clinical related than cell lines for drug screening. Therefore, patient-derived tumor organoids can be used to predict patient responses for novel targeted drugs. However, not for all patients with molecular pathway alterations may be susceptible to molecular targeted treatment. The low success rate of drug testing may be due to the drug was not tested in the proper patient group. Accordingly, to explain which targeted agents correspond which molecular patterns, large studies of prospective biomarkers are necessary (33). For this reason, molecular classification of organoids may help to select the reasonable targeted agents suited for drug testing in organoid models. Moreover, organoids generated from tissue stem cells are mainly structured by the epithelial cells and lack of stromal and immune cell types(10). Thus, transcriptomic profiles represent intrinsic gene patterns of the tumor cells which may enhance molecular subtypes classification based on pathway alterations of CRC organoids.

CHAPTER III

RESEARCH METHODOLOGY

3.1 Data collection

PDOs of colorectal cancer were obtained from Chulalongkorn cancer organoid bank. These PDOs were generated from tissues of stage II and III CRC patients who could be treated with neoadjuvant chemoradiation follow by the ESMO guideline for rectal cancer 2013 or patients with metastatic cancer non-responding to standard treatment**.** The PDOs were maintained according to culture protocol of previous study (35). Samples in this study consists of 55 colorectal cancer PDOs and 5 organoids derived from adjacent normal tissue. RNA was extracted from PDOs using Qiagen kit. mRNA isolation with poly(A) mRNA magnetic isolation module. Then, the libraries were subjected to 2x150bp paired-end sequencing on an Illumina HiSeq instrument. In addition, DNA was extracted from CRC organoids and corresponding peripheral blood mononuclear cells (PBMCs) with a QiaAmp Blood mini kit (Qiagen). For whole exome sequencing, SureSelect Human V6-Post (Agilent), an exome capture kit was used according to the manufacturer's instructions then it was sequenced using Illumina Hiseq2500 (outsourced to Macrogen, Inc.). The quality of sequencing data were visually checked using FastQC (36).

3.2 Somatic variant calling

Whole exome sequencing data in FASTQ format were aligned to the human reference genome version GRCh38 (hg38) using Burrows-Wheeler Aligner (version 0.7.17)(37). Alignment results in SAM format were pre-processed using the Genome Analysis Toolkits (GATK, version 4.1.2.0) (38, 39) according to the best practice developed by the authors. This step removes duplicate reads and recalibrates base calling quality scores (38). Processed whole exome sequencing data from the patient's tumor tissue and PBMC were compared using the Mutect2 module in GATK to identify tumor-specific somatic variants. Mutect2 removes non-tumor-specific variants by comparing variants identified in the tumor sample to those found in the matched PBMC or a panel of normals (PONs), which consists of sequencing data from other healthy individuals. Mutect2 also estimates the extent of contamination of normal cells in tumor sample and utilizes germline allele frequency information from a population of healthy individuals to select tumor-specific somatic variants. Panel of normals and germline allele frequency data were obtained from WES data of peripheral blood. Funcotator was used to annotate the clinical

impact and biological function of each identified variant. Finally, maftools (40) was used to visualize the output.

3.3 Copy number variation (CNV) calling and MSI score

CNVs were called from processed whole exome sequencing data using CNVkit (41). Firstly, “coverage” command computes the log2 mean read depth for a sample using an aligned sequencing reads in BAM format and the target bins in BED format. Then, the “reference” command estimates the expected read depth of each bin across a panel of control samples to produce a reference copy-number profile that can then be used to correct other test samples. Next, the test samples were normalized to the reference using “fix” command. After correction of coverage biases the copy ratio estimates of each sample can be segmented into dicrete copy-number regions using the “segment” command. Finally, log2 copy ratio of multiple samples were visualized as a heatmap. For MSI score, paired tumor-normal whole exome sequencing data of each CRC organoid were investigated MSI sites through MSIsensor followed by recommended pipeline (42).

3.4 Differential gene expression analysis

RNA sequencing data were first trimmed using Cutadapt (v1.9.1) (43) and subsequently aligned to human reference genome (hg38) and quantified using Kallisto (version 0.46.2) (44) with 20 bootstraps. Next, differential gene expression between normal and cancer organoids were analyzed through Sleuth (version 0.30.0) (45) R package. Differentially expressed genes (DEGs) were reported (adjusted p-value < 0.05). Gene expression level (tpm) were presented by boxplot through ‘ggplot2’ R package (46).

3.5 unsupervised clustering of gene expression data

Hierarchical clustering was performed using the pvclust R package (version 2.2.0) (47). Expression data of 747 DEGs between cancer and normal organoids were normalized to count per million (CPM), loP2 transformed, and used as input for the clustering. The hierarchical clustering process was repeated with 10,000 bootstraps to assess the uncertainty. Pearson correlation distance and average linkage method were selected.

Non-negative matrix factorization (NMF) was performed using the NMF R package (48). The expression matrix of the 12,529 high variance (variance >1) genes was analyzed to identify the predetermined number of clusters (K) varying from 2 to 6. At each number of cluster setting, 40 iterations of NMF were performed. The algorithm of Brunet et al. (49) was selected. Quality of

the clustering was evaluated by the cophenetic coefficient and the number of clusters at which the coefficient began to drop was chosen as the optimal number of clusters (31).

3.6 Functional annotation and enrichment analysis

To characterize the gene functional signature of each PDO cluster, differential expression analysis was performed between PDOs in that cluster against all other cancer PDOs as described above. DEGs (adjusted p-value < 0.05) and their corresponding log2 fold difference values were than submitted to a gene set enrichment analysis (GSEA) (50) against the KEGG pathway databases using the WebGestalt interface (http://www.webgestalt.org/) (51). Top 10 up- and down-regulated pathways are listed in Table1.

3.7 Consensus molecular subtype (CMS) and CRC intrinsic subtype (CRIS)

CMS subtypes for PDOs in this study were predicted using DeepCC (52). For DeepCC, the log2 transformed expression data of CRC organoids were used as input. Additionally, CRC intrinsic subtype (CRIS) prediction was performed through the CRISclassifier R package (9). All prediction results were filtered using an adjust p-values ≤ 0.05.

3.8 prediction of subtype clusters from unsupervised analysis

Transcriptomic data of 14283 genes of 54 samples of stage II and III CRC were first normalized and log transformed. The processed expression data were then divided into a training and a test dataset with 35 and 19 samples, respectively. The least absolute shrinkage and selection operator (LASSO) logistic regression models were trained using the glmnet R package (53) to classify each sample according to the clusters identified via NMF method. 3-fold cross validation was performed on the training dataset to tune the regularization parameter λ of the LASSO model. The value of λ that yielded the lowest average classification error over cross-validation was selected. Finally, a LASSO logistic regression model was trained using the whole training set and evaluated using the test dataset. Genes with nonzero coefficients in this model were designated as signature genes for the NMF clusters.

3.9 co-expression matrix

Gene expression data of CRC organoids in ribosome biogenesis pathway were extracted using gene list from the Molecular Signatures Database (MSigDB). Next, these genes were calculated for Pearson correlation then plotted in heatmap using pheatmap R package (54). Hierarchical clustering of genes was performed using Euclidean distance and average clustering method. Clusters were identified by using cutree R function.

CHAPTER IV

RESULTS

4.1 Genetic profiling of individual CRC organoids

To investigate genomic alterations of CRC organoids 38 whole-exome sequencing data were examined. Firstly, according to the majority of CRC demonstrated chromosomal instability (CIN) during cell division and this feature leads to gains and losses of various genes thus copy-number variation (CNV) was analyzed. The results demonstrated that 18 out of 38 organoid samples (47.37%) present high CNV and these samples shown deletion of chromosome 18 involving several tumor suppressor genes (Figure1). This result indicated the chromosomal instability (CIN) feature of organoids in this study. Secondly, the most frequently mutated genes of CRC were found in these organoids including APC, TP53 and KRAS genes (Figure3A). Then, the mutation frequency of organoids was compared to TCGA database (Figure3B). The CRC organoids in this study exhibit slightly different mutation frequencies of these genes. Nevertheless, organoids presented lower frequency in important genes than that were found in the TCGA database such as APC, TP53 and KRAS mutations. Lastly, to evaluate the microsatellite instability (MSI) status of all organoids, MSI scores were calculated. The results revealed that 4 out of 38 CRC organoids (10.53%) have high MSI scores referring to abnormalities of DNA mismatch repair (MMR) (Figure2). The CRC organoids with MSI high have the possibility of increased gene mutation leading to distinct biologic characteristics compared to the microsatellite stable group. In summary, these genomic profiling results demonstrated that CRC organoids generated in this study contained diverse genetic alterations and consistent with that were found in colorectal cancer.

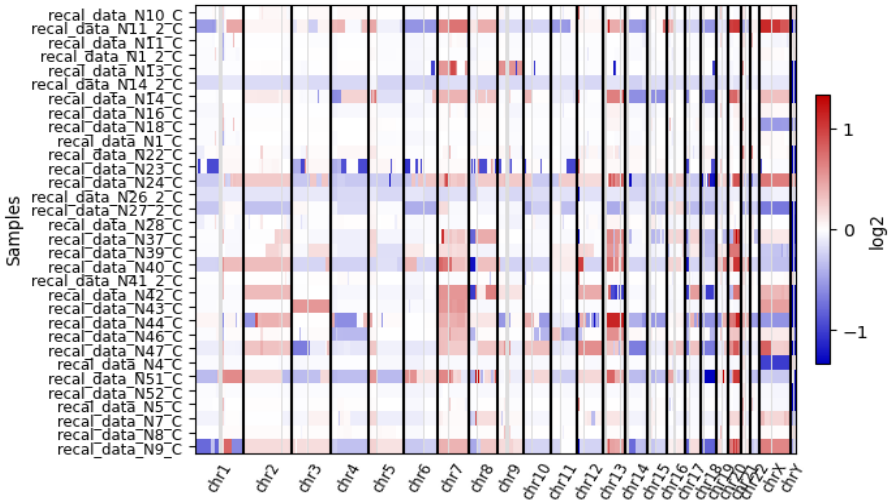


Figure 1 Copy-number variations (CNV) heatmap of organoids

heatmap visualize log2 gene copy ratio of each sample red and blue indicate amplification and deletion of copy number, respectively.

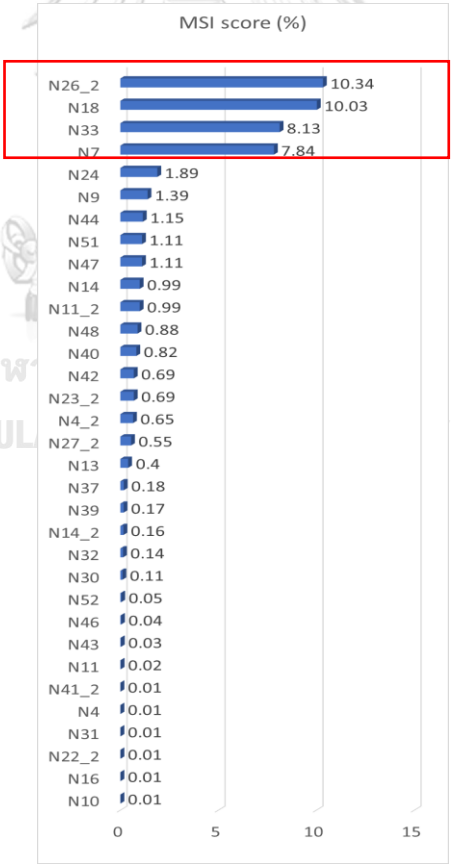
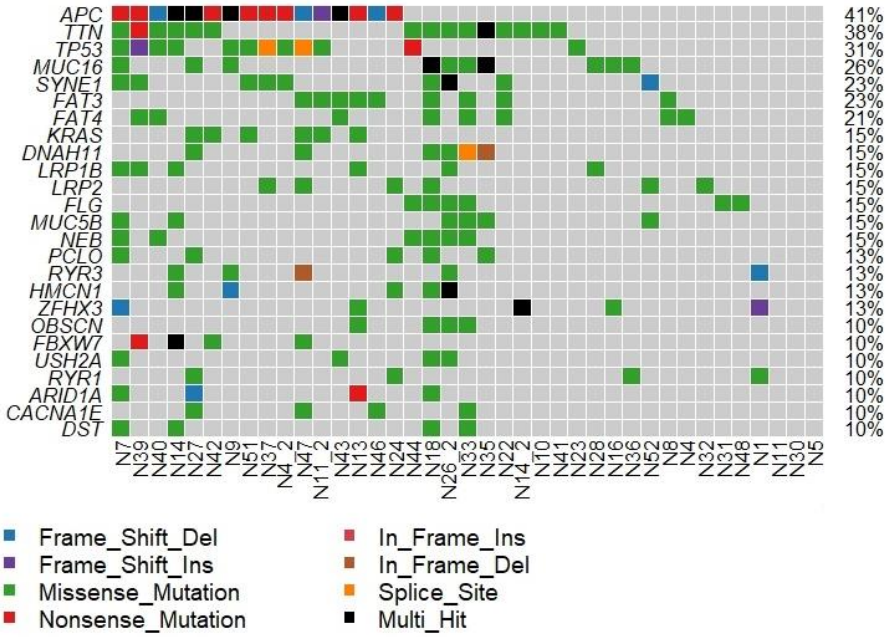


Figure 2 MSI score (%) of individual CRC organoid

A



B

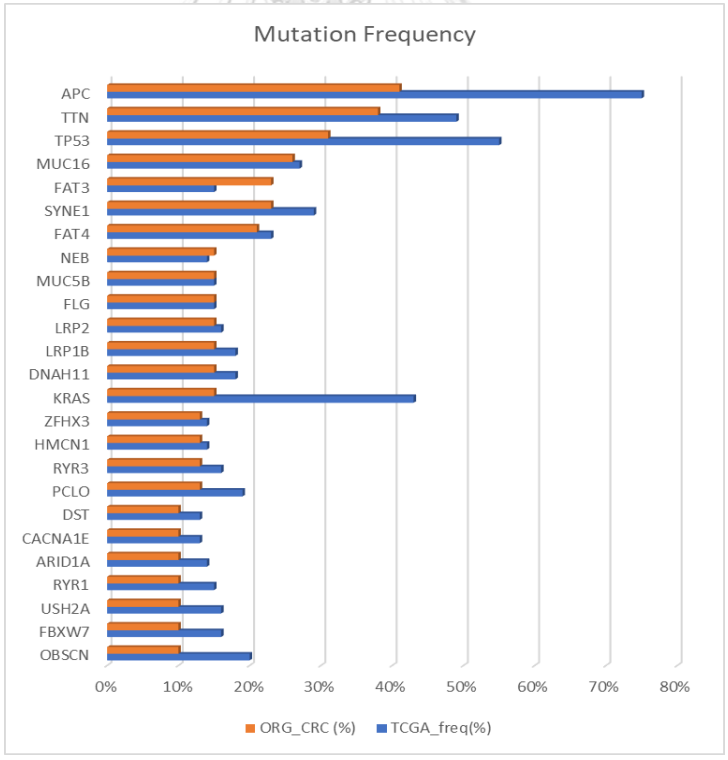


Figure 3 Mutation patterns of 38 CRC organoids in this study

(A) Somatic mutations of CRC frequently mutated genes; (B) mutation frequency of CRC organoids (orange) and TCGA database (blue)

4.2 Clustering of CRC organoids transcriptomic data

To classify molecular subtypes of CRC organoids, unsupervised clustering of transcriptomic data from 54 CRC organoids with non-negative matrix factorization (NMF) method was analyzed. The clustering results indicated that the data can be robustly delineated into groups at various resolutions, from k = 2 to k = 6 clusters (Figure 4A). Using cophenetic coefficient to measure the quality of clustering revealed that the optimal number of clusters is at k = 4 after which the coefficient steadily drops (Figure4B). With k =4, the clusters classified by NMF are extremely robust (cophenetic coefficient = 0.996). Thus, using this method can classify CRC organoids into four groups: P1 (14/54; 25.93%), P2 (15/54; 27.78%), P3 (12/54; 22.22%) and P4 (13/54; 24.07%).

As an alternative, hierarchical clustering (pvclust method) was also performed on the transcriptomic data. Here, instead of considering all genes, a set of 747 genes that are differentially expressed between cancer and normal organoids were selected. Transcriptomic data from 5 paired normal organoids were also included in the analysis. The dendrogram was then constructed with average linkage and correlation distance (Figure 5A). This shows a clear separation between normal and CRC organoids and suggests that CRC organoids may be classified into up to four groups (G1-G4). While G1 and G2 are well-separated, there is no clear boundary between G3 and G4.

Comparison between the clusters identified by pvclust and NMF indicate a good agreement, especially between G1 and G4 groups of pvclust method and P1 and P4 groups of NMF (Figure 5B). As the NMF method was more objectively tuned using cophenetic coefficient and did not rely on gene selection, the clusters identified by NMF (P1-P4) were selected to further analyses.

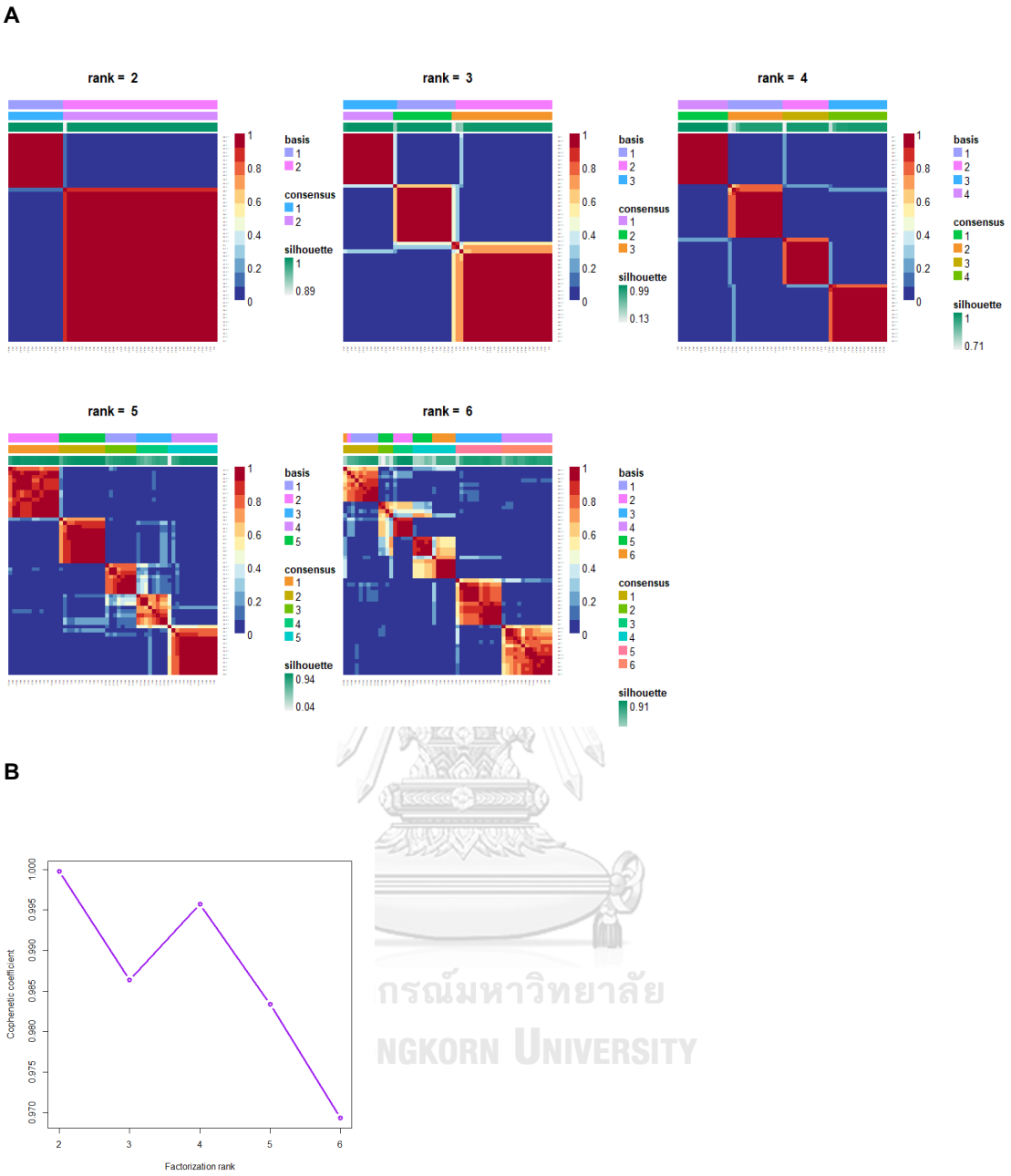


Figure 4 NMF-based consensus clustering of the PDOs transcriptomic data

(A) Consensus map of NMF clustering result (k = 2-6) (B) The trend of cophenetic coefficient as the number of cluster (k) increases. k=4 is selected as the optimal number of cluster as it is where the cophenetic coefficient begins to drop

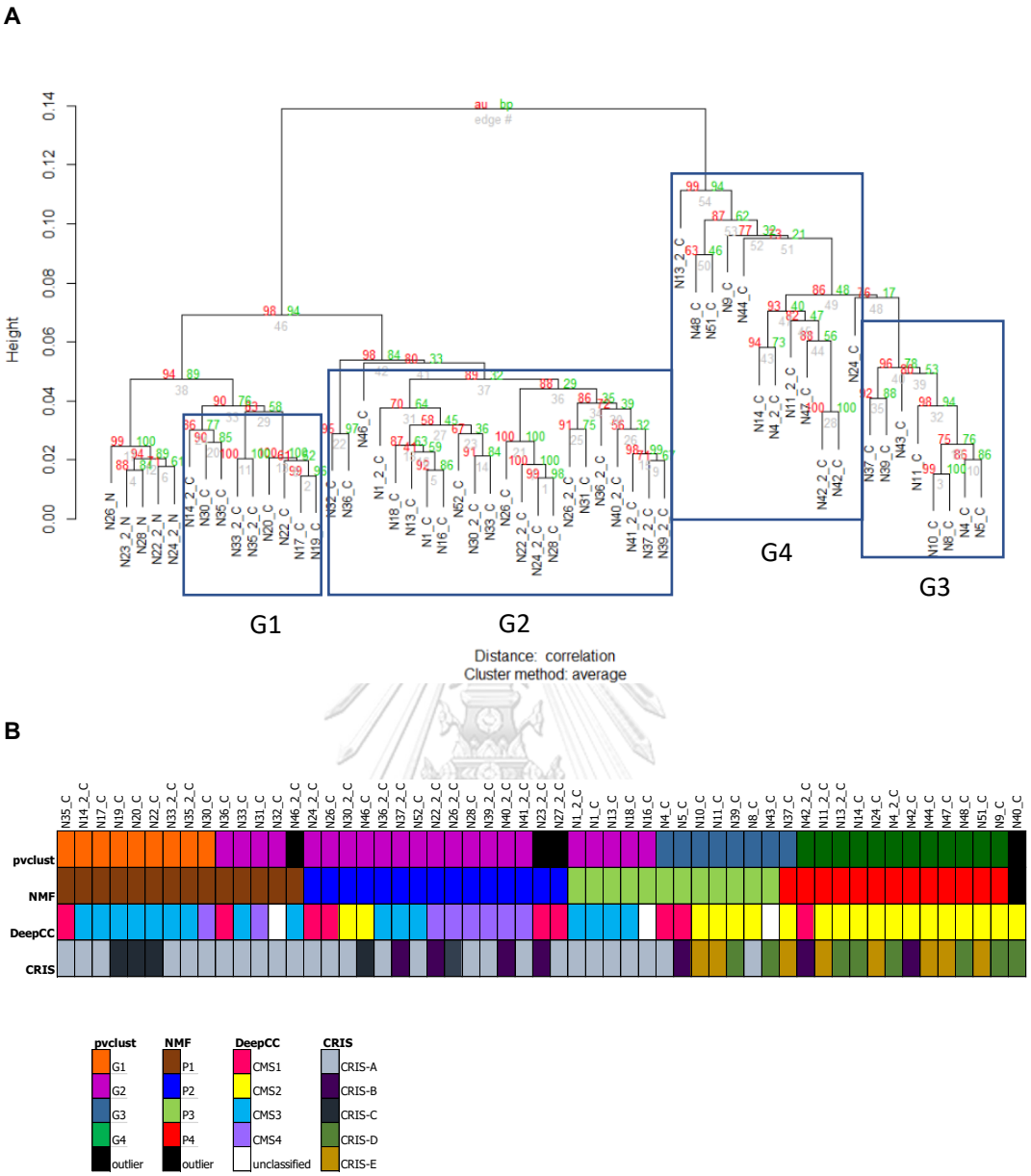


Figure **5** *Hierarchical clustering of PDOs transcriptomic data based on 747 differentially expressed genes.*

(A) Hierarchical clustering results from pvclust R package. Approximately Unbiased p-values (red) and Bootstrap p-values (green) produced by pvclust were shown at each branching point. (B) Comparison of predicted CMS subtype, predicted CRIS subtype, and clustering results obtained from pvclust or NMF.

4.3 comparison between NMF-derived clusters, CMS, and CRIS

To compare NMF-derived clusters with the established CMS and CRIS systems, the subtypes of CRC organoids were also predicted using DeepCC (52) and CRISclassifier (9), respectively. The results shown in Figure5B demonstrate that most samples in P1 were predicted as CMS3 (metabolic subtype), CRIS-A (hypoxic and glycolytic subtype), and CRIS-C (Epidermal growth factor receptor (EGFR) pathway activation). P2 and P3 are associated with multiple CMS and CRIS subtypes. Interestingly, CMS4 was exclusively predicted only in P2 and shared a similar characteristic of TGFβ upregulation with CRIS-B by. Moreover, CRIS-D and CRIS-E, which are associated with upregulation in WNT pathway, were identified in P4 which is predominantly predicted as CMS2 (canonical wnt subtype).

4.4 Functional characteristic of individual group

To better describe the functional characteristics inherent to each CRC organoid group defined by NMF clustering, differential expression analysis was performed to identify significantly up- or down-regulated genes in that group compared to the others. Next, Gene Set Enrichment Analysis (GSEA) was applied to identify enriched pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. The top 10 of KEGG pathways associated with each cluster were listed in Table2 and described in more details below.

Table 2 The results of top 10 up- and down-regulated KEGG pathway enrichment influenced by the differentially expressed genes of individual group.

Group	Up-regulated				Down-regulated			
	Gene Set	Description	P Value	FDR	Gene Set	Description	P Value	FDR
P1	KEGG: hsa00140	Steroid hormone biosynthesis	0.000 *	0.000 *	KEGG: hsa03030	DNA replication	0.000 *	0.000 *
	KEGG: hsa00830	Retinol metabolism	0.000 *	0.000 *	KEGG: hsa03010	Ribosome	0.000 *	0.000 *
	KEGG: hsa04975	Fat digestion and absorption	0.000 *	0.001 *	KEGG: hsa04110	Cell cycle	0.000 *	0.000 *
	KEGG: hsa00040	Pentose and glucuronate interconversions	0.000 *	0.002 *	KEGG: hsa03460	Fanconi anemia pathway	0.000 *	0.005 *
	KEGG: hsa04144	Endocytosis	0.000 *	0.003 *	KEGG: hsa03008	Ribosome biogenesis in eukaryotes	0.002 *	0.012 *
	KEGG: hsa00982	Drug metabolism	0.003 *	0.003 *	KEGG: hsa03013	RNA transport	0.000 *	0.012 *
	KEGG: hsa00601	Glycosphingolipid biosynthesis	0.000 *	0.004 *	KEGG: hsa03440	Homologous recombination	0.000 *	0.013 *
	KEGG: hsa04923	Regulation of lipolysis in adipocytes	0.002 *	0.007 *	KEGG: hsa03420	Nucleotide excision repair	0.003 *	0.047 *
	KEGG: hsa05204	Chemical carcinogenesis	0.003 *	0.007 *	KEGG: hsa04310	Wnt signaling pathway	0.002 *	0.049 *
	KEGG: hsa04972	Pancreatic secretion	0.000 *	0.014 *	KEGG: hsa03410	Base excision repair	0.007 *	0.049 *
P2	KEGG: hsa04310	Wnt signaling pathway	0.007 *	0.085	KEGG: hsa05169	Epstein-Barr virus infection	0.879	0.874
	KEGG: hsa05205	Proteoglycans in cancer	0.003 *	0.097	KEGG: hsa05222	Small cell lung cancer	0.853	0.966
	KEGG: hsa05225	Hepatocellular carcinoma	0.030 *	0.286	KEGG: hsa05203	Viral carcinogenesis	0.174	0.987
	KEGG: hsa05165	Human papillomavirus infection	0.206	0.534	KEGG: hsa01100	Metabolic pathways	0.128	1.000
	KEGG: hsa04060	Cytokine-cytokine receptor interaction	0.066	0.536	KEGG: hsa04714	Thermogenesis	0.367	1.000
	KEGG: hsa05146	Amoebiasis	0.207	0.546	KEGG: hsa04024	cAMP signaling pathway	0.530	1.000
	KEGG: hsa04010	MAPK signaling pathway	0.093	0.557	KEGG: hsa04723	Retrograde endocannabinoid signaling	0.642	1.000
	KEGG: hsa04933	AGE-RAGE signaling pathway in diabetic complications	0.199	0.557	KEGG: hsa04380	Osteoclast differentiation	0.665	1.000
	KEGG: hsa04144	Endocytosis	0.271	0.565	KEGG: hsa04520	Adherens junction	0.759	1.000
	KEGG: hsa04390	Hippo signaling pathway	0.111	0.572	KEGG: hsa05202	Transcriptional misregulation in cancer	0.866	1.000
P3	KEGG: hsa03010	Ribosome	0.000 *	0.149	KEGG: hsa04144	Endocytosis	0.005 *	0.172
	KEGG: hsa04110	Cell cycle	0.006 *	0.236	KEGG: hsa04270	Vascular smooth muscle contraction	0.016 *	0.179
	KEGG: hsa04360	Axon guidance	0.011 *	0.254	KEGG: hsa04923	Regulation of lipolysis in adipocytes	0.007 *	0.186
	KEGG: hsa04060	Cytokine-cytokine receptor interaction	0.008 *	0.315	KEGG: hsa00561	Glycerolipid metabolism	0.018 *	0.193
	KEGG: hsa05323	Rheumatoid arthritis	0.010 *	0.351	KEGG: hsa04142	Lysosome	0.021 *	0.193
	KEGG: hsa04115	p53 signaling pathway	0.044 *	0.456	KEGG: hsa04915	Estrogen signaling pathway	0.028 *	0.236
	KEGG: hsa04064	NF-kappa B signaling pathway	0.038 *	0.473	KEGG: hsa02010	ABC transporters	0.004 *	0.242
	KEGG: hsa04390	Hippo signaling pathway	0.041 *	0.493	KEGG: hsa04213	Longevity regulating pathway	0.057	0.244
	KEGG: hsa04062	Chemokine signaling pathway	0.037 *	0.523	KEGG: hsa04611	Platelet activation	0.070	0.245
	KEGG: hsa04145	Phagosome	0.046 *	0.566	KEGG: hsa04970	Salivary secretion	0.066	0.258
P4	KEGG: hsa03030	DNA replication	0.000 *	0.000 *	KEGG: hsa05146	Amoebiasis	0.000 *	0.001 *
	KEGG: hsa03460	Fanconi anemia pathway	0.000 *	0.000 *	KEGG: hsa00982	Drug metabolism	0.000 *	0.002 *
	KEGG: hsa03008	Ribosome biogenesis in eukaryotes	0.000 *	0.008 *	KEGG: hsa04144	Endocytosis	0.000 *	0.003 *
	KEGG: hsa03440	Homologous recombination	0.000 *	0.009 *	KEGG: hsa04060	Cytokine-cytokine receptor interaction	0.000 *	0.008 *
	KEGG: hsa05033	Nicotine addiction	0.002 *	0.032 *	KEGG: hsa05418	Fluid shear stress and atherosclerosis	0.000 *	0.019 *
	KEGG: hsa04110	Cell cycle	0.000 *	0.035 *	KEGG: hsa00512	Mucin type O-glycan biosynthesis	0.004 *	0.019 *
	KEGG: hsa03430	Mismatch repair	0.010 *	0.065	KEGG: hsa04510	Focal adhesion	0.000 *	0.020 *
	KEGG: hsa03013	RNA transport	0.004 *	0.071	KEGG: hsa00040	Pentose and glucuronate interconversions	0.000 *	0.020 *
	KEGG: hsa03410	Base excision repair	0.025 *	0.098	KEGG: hsa04668	TNF signaling pathway	0.000 *	0.023 *
	KEGG: hsa03020	RNA polymerase	0.015 *	0.105	KEGG: hsa04810	Regulation of actin cytoskeleton	0.000 *	0.025 *

4.4.1 P1: metabolic pathway alterations

GSEA-based phenotypic analyses reveal metabolic pathway alteration of P1. It is enriched pathway involved in metabolism of lipid and cholesterol including fat digestion and absorption, regulation of lipolysis in adipocytes, retinol metabolism and steroid hormone biosynthesis pathways (Table2). Moreover, organoids in P1 also show upregulated pentose and glucuronate interconversions pathway. As expected, by DeepCC, 9 out of 14 samples in P1 were predicted as CMS3 metabolic subtype (Figure5B). Then, the enriched pathways of P1 were compared to CMS3. The result indicated that CMS3 presented alterations in diverse metabolic pathways such as glucose and pentose metabolism, nitrogen metabolism and fatty acid metabolism etc. Interestingly, CMS3 were enriched in metabolic of phospholipid and fatty acids which might be associated with lipid metabolism of P1.

Then, to explore additional characteristic of P1, Crypt top and crypt base gene signatures of colon from previous study (55) were applied to investigate expression pattern of each groups. Interestingly, P1 organoids expressed higher signature of crypt top signature indicated that it presented more kind of differentiated cell than others (Figure10A).

4.4.2 P2: highly expressed WNT7A and WNT7B and TGFβ activation

Comparison between P2 and other groups, 379 DEGs were identified before applied these DEGs as input of GSEA. However, KEGG pathway enrichment results of P2 presented three pathways were significantly enriched (p value < 0.05) including Wnt signaling, proteoglycans in cancer, and Hepatocellular carcinoma pathways (Table2). Wnt signaling pathway which is one of the most frequent abnormalities in human cancer. When explore the wnt signature the results demonstrated that P2 was not presented the highest wnt pathway (Figure6D). However, P2 show higher expression of WNT7A and WNT7B which are ligand of this pathway than other groups (Figure6A-C). Furthermore, TGFB1 ligand and SMAD3 downstream target of TGF beta pathway significantly upregulated in P2. Additionally, this groups also show higher expression of mesenchymal signature than other groups. Importantly, 3 out of 7 samples of P2 contain mutation in SMAD4 gene which is downstream target gene of this pathway. However, organoid culture media contain TGF beta inhibitor which might result in ambiguous difference between P2 and P3 thus withdrawal of grow factors from culture media is needed to further investigate molecular characteristic of these groups.

For DeepCC results, 6 out of 15 samples in P2 were predicted in CMS4 which is mesenchymal subtype and highly expressed TGF beta pathway (Figure5B). Together with molecular characteristics of P2, these results indicate the association between P2 and CMS4.

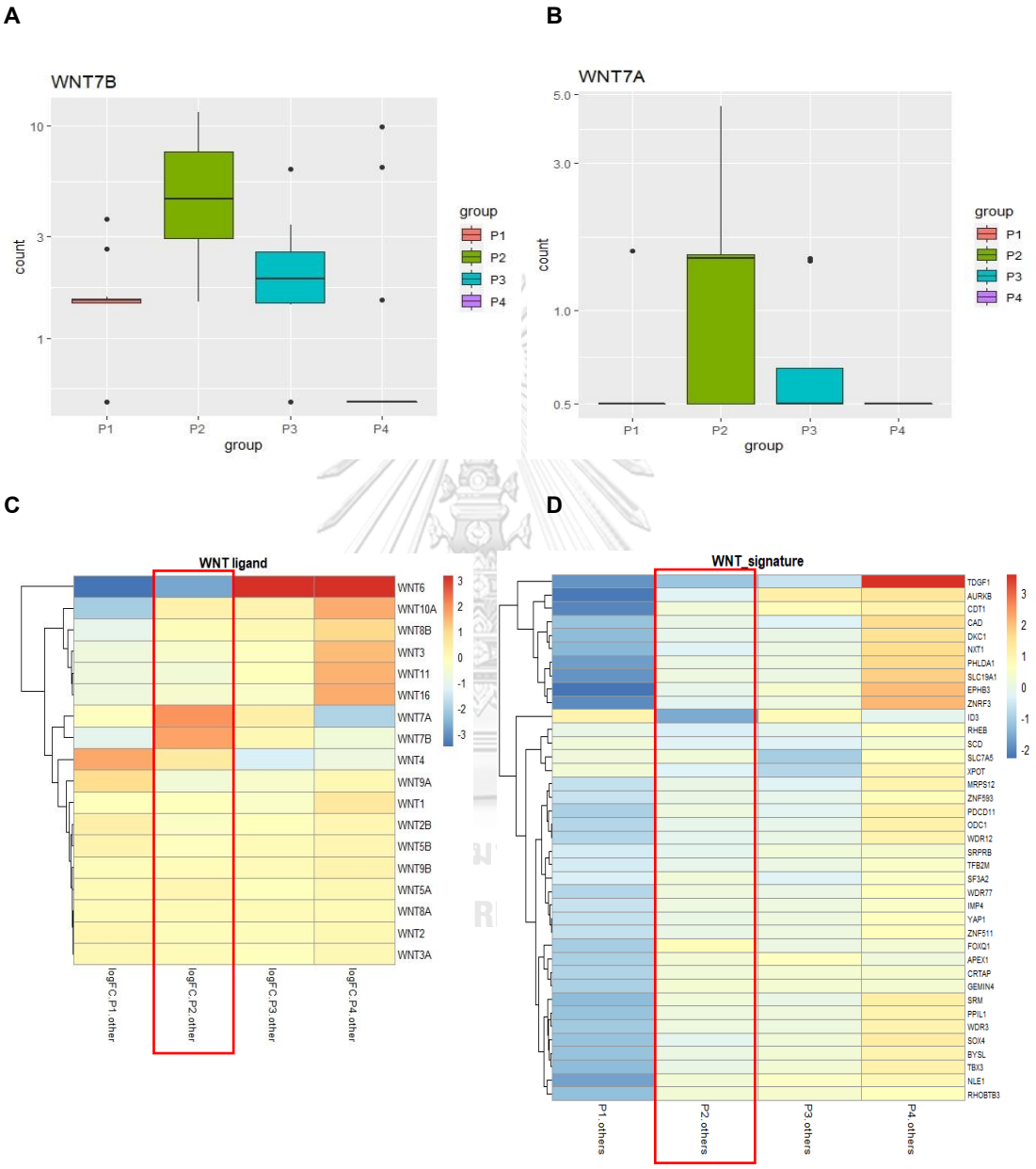


Figure 6 Expression levels of wnt signaling pathway in P2

(A) WNT7B, (B) WNT7A, (C) heatmap of WNT ligand expression and (D) heatmap of WNT signature expression

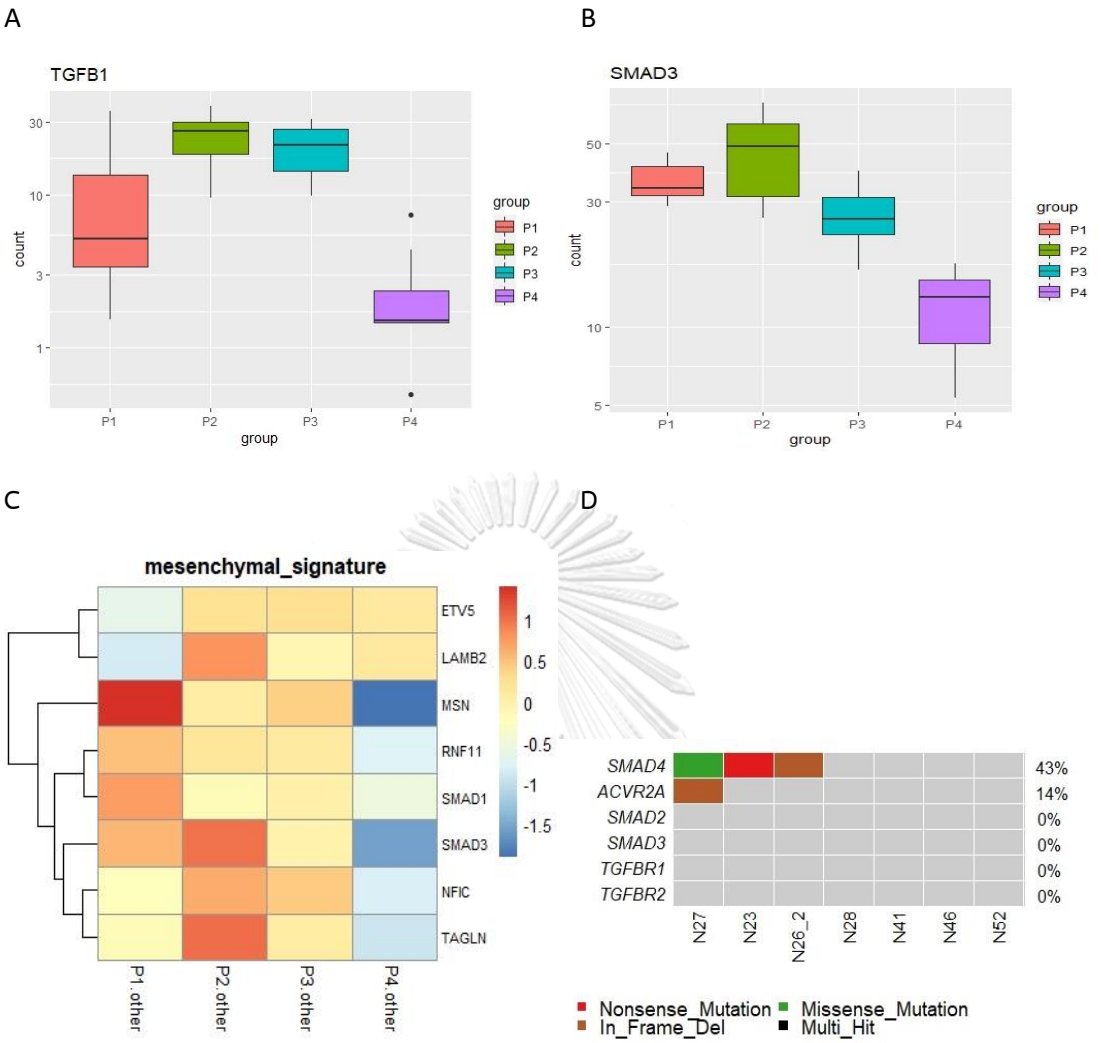


Figure 7 Expression levels of TGF-β pathway in P2

(A) TGFB1 (B) SMAD3 (C) mesenchymal signature (D) mutations in TGFβ pathway **of P2 organoids**

4.4.3 P3: TGFβ and chemokines activation

After performed functional enrichment analysis to 1173 DEGs of P3, the results indicated that these pathways are not significantly enriched with False discovery rate (FDR) < 0.05 (Table2). However, associated genes of P3 were further identified including chemokine-related genes such as CXCL8 and CCL2 (Figure8B, C). Furthermore, BCL2, apoptosis suppressor gene, demonstrated significantly high expression in P3 compared to other groups (Figure8D). Additionally, AQP1 and TGFBI involving in cell migration demonstrated significantly high expression in this group (Figure8E, F). Besides, organoids in P3 presented high expression of TGFB1 and TGFβ pathway but it slightly lower than P2 (Figure8A). According to this group showed some similar characteristic with P2 thus growth factors withdrawal which was previously mentioned might allow to separate difference between these two groups. Finally, CMS prediction results of DeepCC in P3 showed diverse subtypes including CMS1-3 and unclassified subtype (Figure5B).

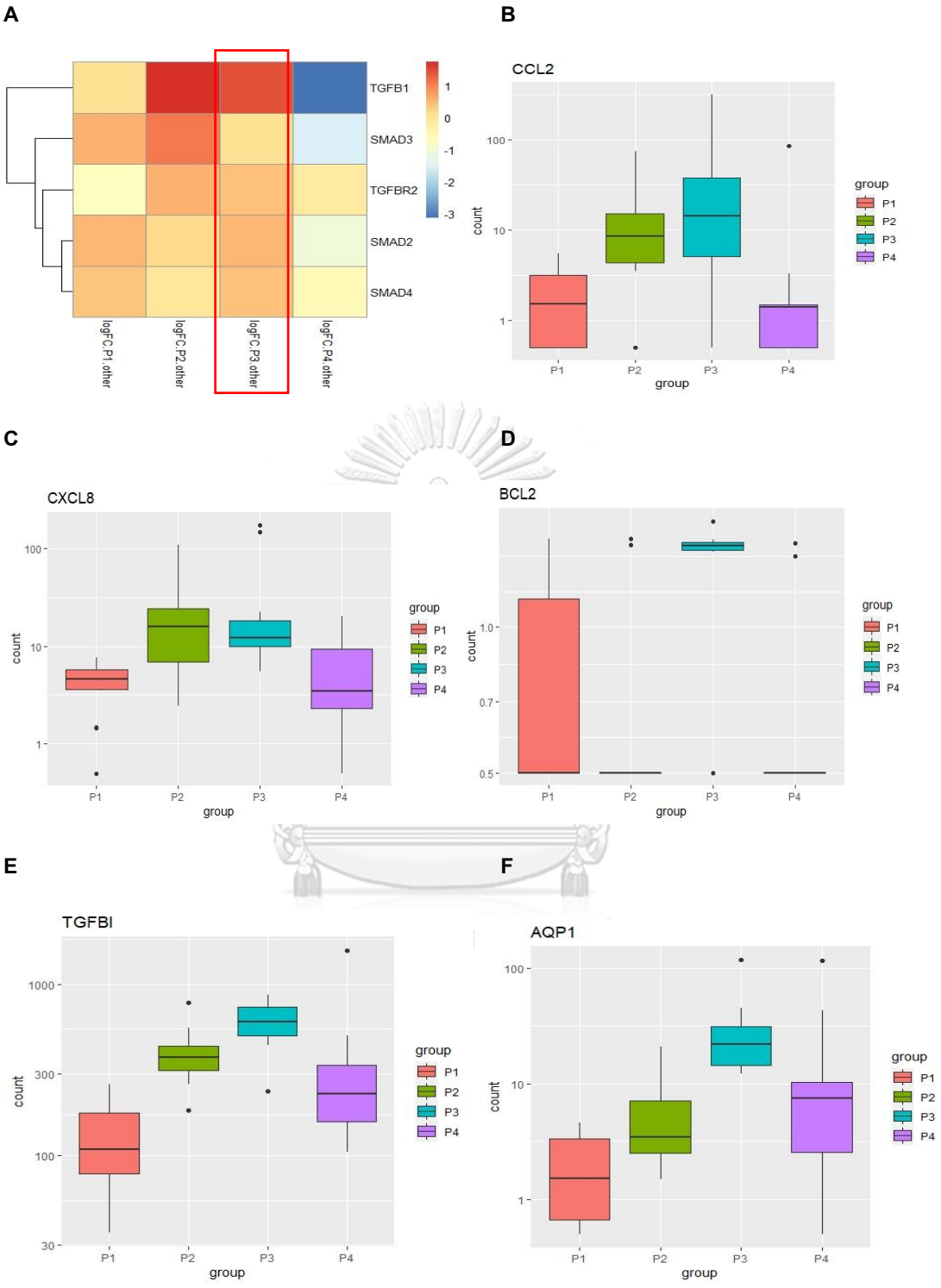


Figure 8 Highly expressed genes in P3

(A) Heatmap of TGFβ pathway (B-F) Expression levels of CCL2, CXCL8, BCL2, TGFBI and AQP1 genes, respectively.

4.4.4 P4: stem cell-like features and DNA damage responses associated with drug resistance

In P4 10,467 DEGs were applied to GSEA then the top10 significantly up- and down-regulated pathways were shown in Table2. The results demonstrated that organoid in P4 are highly expressed genes involved in DNA-damaged repair pathway including base excision repair, Fanconi anemia (FA) pathway, homologous recombination (HR), and mismatch repair (MMR) pathway. These pathways related to DNA damaged response (DDR) which plays an important role in the maintenance of genome stability and integrity through correcting the impaired DNA that may contribute to carcinogenesis (56). Importantly, when combined with radiation response of organoid from previous study P4 organoids presented radiation resistance which reveal aggressive tumor of this group (Figure9). Thus, we hypothesized that upregulated DDR pathway associated with radiation resistance by relieving DNA lesions and chromosomal abnormalities that occur together with cancer cell proliferation. Interestingly, this molecular feature was found in cancer stem cells (CSCs) ae well to resist DNA damage repair capacity and protecting DNA damage by an efficient scavenging of reactive oxygen species (ROS), generated by the chemotherapy or radiotherapy.

Then, to investigate stem cell features of P4 organoids, gene signature of LRG5+ intestinal stem cells (ISCs) and WNT expression from previous study were applied (57). The results were presented in heatmaps (Figure10C, D). Unsurprisingly, P4 highly expressed gene signature of colon crypt base which intestinal stem and progenitor cells are located (Figure10B). This result agrees with upregulation of LGR5+ ISC and Wnt signatures. Hence, we hypothesized that organoids in P4 have stem cell-like patterns.

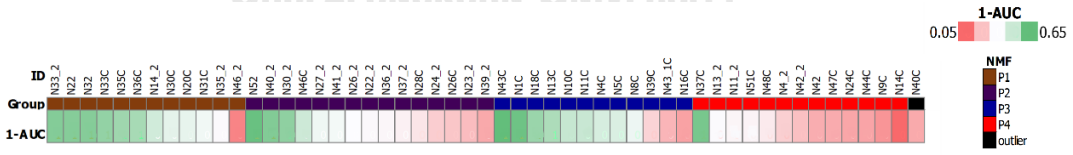


Figure ***9 Radiation response of CRC organoids***

The heatmap presents Area under the survival curve (AUC) after CRC organoids exposed different doses of radiation

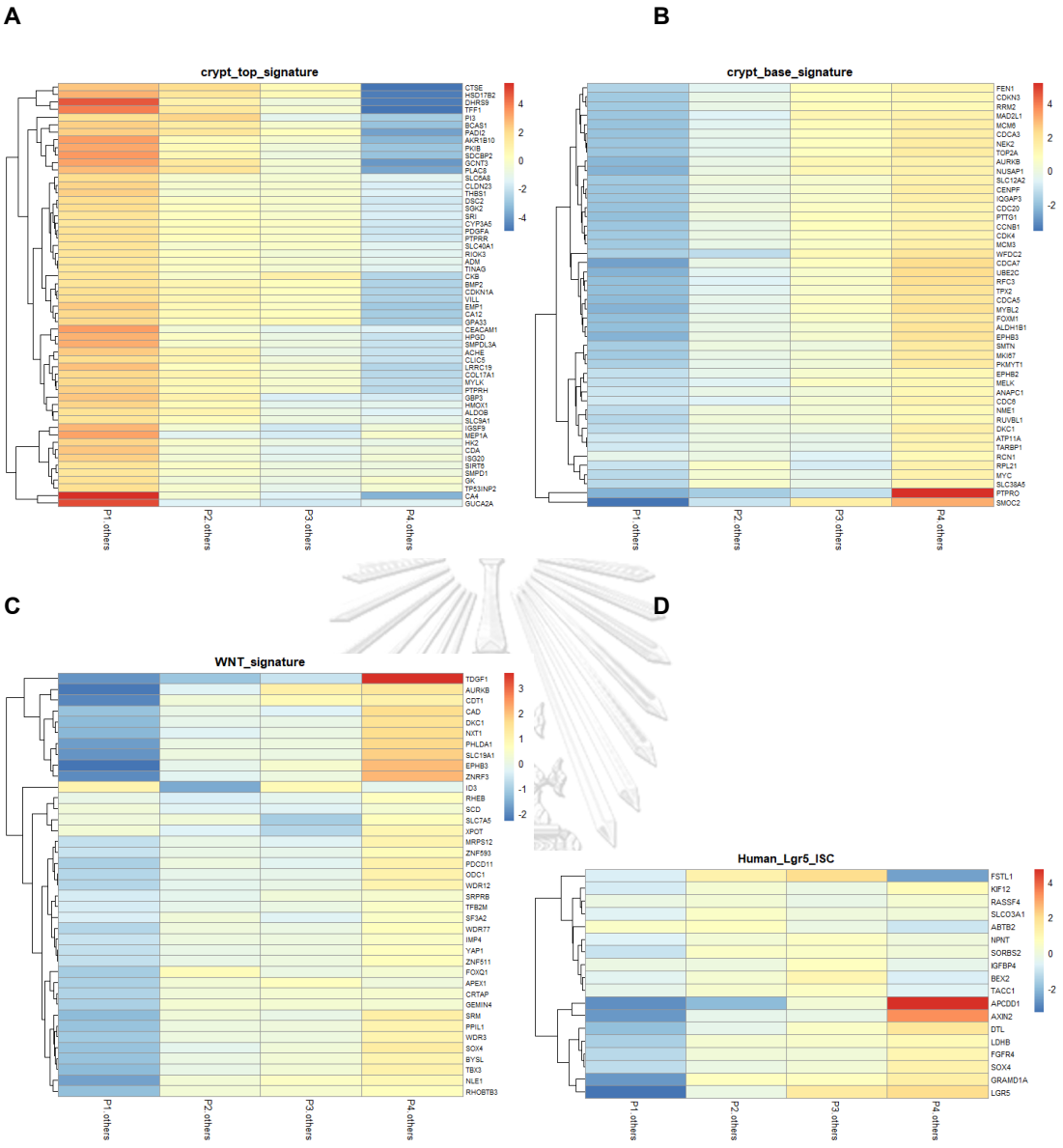


Figure **10** *Stem cell-like properties of P4*

(A) crypt top signature, (B) crypt base signature, (C) gene in WNT signaling pathway and (D) human Lgr5+ intestinal stem cells (ISCs). Color in heatmap represented fold changes, the higher expression (red) and lower expression (blue).

4.4.5 P4: Upregulation of ribosome biogenesis pathway

In P4, ribosome biogenesis in eukaryotes, RNA polymerase and RNA transport pathways were highly enriched. These pathways involved in protein synthesis which support to continuous growth of cancer cells (58) and associated with the high activation of cell cycle and DNA replication pathways found in P4. Moreover, the activation of ribosome biogenesis has been comprehensively linked to sustained RNA polymerase I and III activation. Thus, high expression level of RNA polymerase I (POLR1A) and III (POLR3A) of P4 were confirmed in Figure11A, B. Moreover, a prominent role in the regulation of rRNA transcription in cancer is played by the C-MYC proto-oncogene. C-MYC boosts all steps of rRNA biosynthesis and maturation through diverse molecular mechanisms. As expected, organoids in P4 showed significantly high expression of MYC gene (Figure11C). These results supported the remarkable increase of ribosome biogenesis in P4 which might be applied as a biomarker of this group.

Interestingly, mutations in APC, TP53 and KRAS genes were dominantly found in P4 organoids (Figure11D). In addition to these mutations, in P4 organoid group copy-number variations were higher than other groups (Figure11E, F). these results support the tumor aggressiveness of this group.

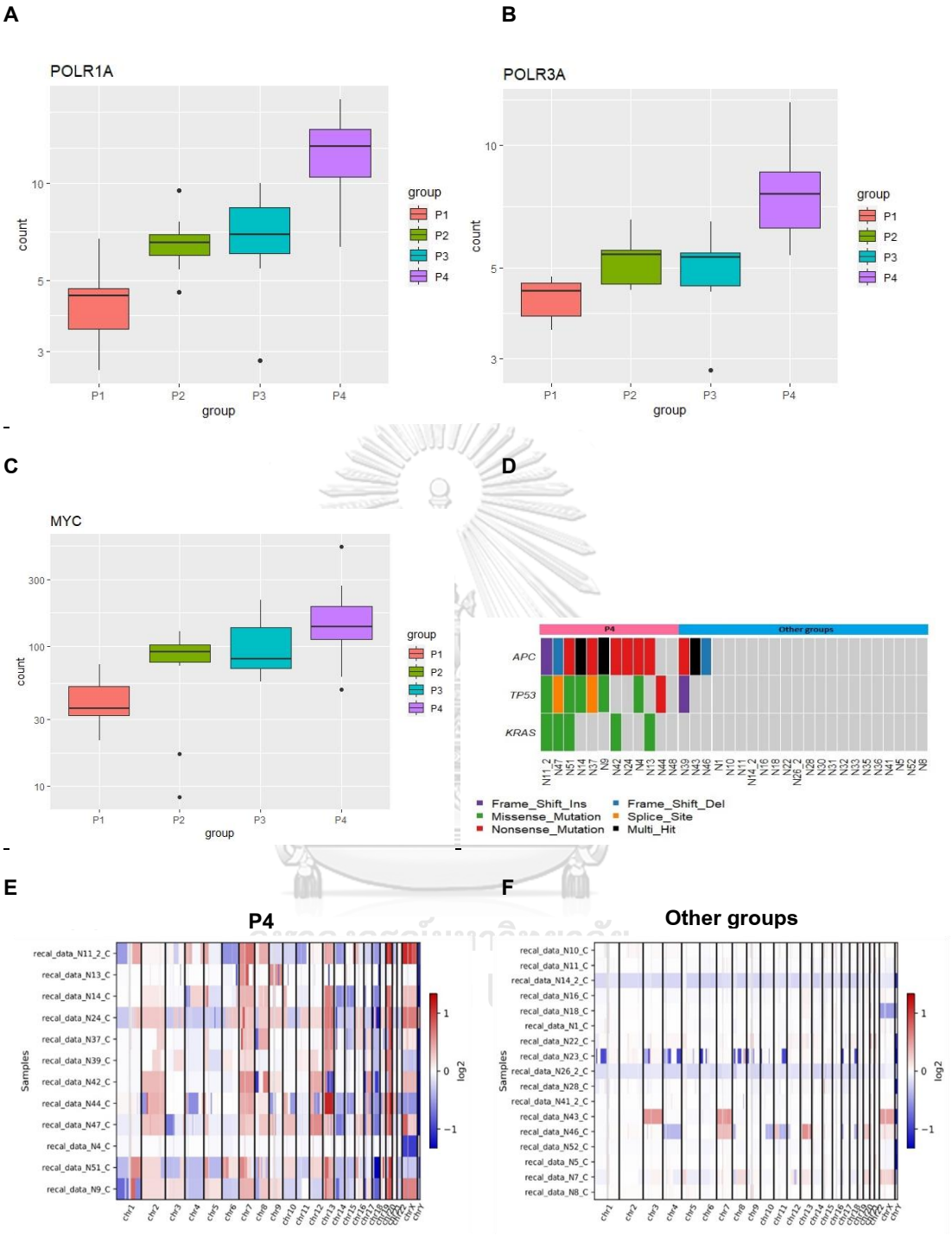


Figure **11** *Expression levels of ribosome biogenesis related genes and genomic profiles of P4* (A) POLR1A, (B) POLR3A, (C) MYC, (D) Mutations of APC, TP53 and KRAS genes found in P4. (E,F) high CNV of P4 compared to other groups.

4.5 Co-expression of genes in Ribosome biogenesis pathway

Transcriptome and clinical data demonstrated that P4 organoids show characteristics of aggressive tumor and ribosome biogenesis pathway was specifically enriched in this group thus we hypothesized that organoids in P4 might be identified by using a few gene in ribosome biogenesis pathway. Hence, to investigate the most correlated gene in this pathway ribosomal biogenesis genes were analyzed as a co-expression matrix. The result demonstrated that in ribosome biogenesis pathway there are six groups of genes functioning as shown in red diagonal (Figure12). Then, overlapped DEGs between P4 and CMS2 were labeled, and the result indicated that mainly of these genes are located in cluster 1 and 2. Consequently, the top10 genes of these two clusters were proposed as genes markers of P4 ribosome biogenesis group (Table3).

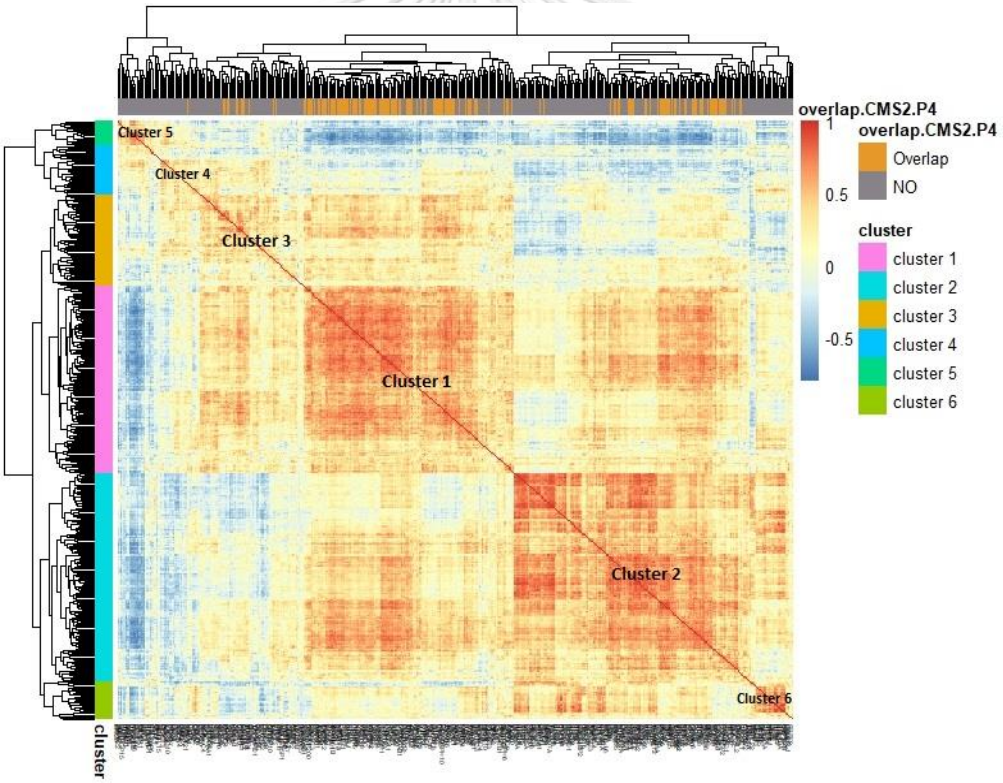


Figure 12 co-expression heatmap of ribosome biogenesis pathway in CRC organoids

Table 3 Top10 genes of cluster 1 and 2 of ribosome biogenesis pathway ranked by log2 fold change of gene expression

Top10 genes of cluster1

SYMBOL	genes	logFC	AveExpr	adj.P.Val
SUV39H1	ENSG00000101945	1.476134	3.52651	4.96E-14
GTF3A	ENSG00000122034	1.300063	6.491799	9.04E-11
UTP14A	ENSG00000156697	1.24507	5.299221	5.22E-13
NSUN5P1	ENSG00000223705	1.230734	4.731074	2.12E-07
DKC1	ENSG00000130826	1.154322	6.651386	5.49E-12
LAS1L	ENSG00000001497	1.114469	5.699689	5.80E-11
PRKDC	ENSG00000253729	1.091219	6.446777	1.65E-09
MTG2	ENSG00000101181	1.011972	5.531135	8.42E-11
LYAR	ENSG00000145220	0.999265	4.093263	3.50E-10
RRP12	ENSG00000052749	0.989937	5.516368	2.22E-09

Top10 genes of cluster2

SYMBOL	genes	logFC	AveExpr	adj.P.Val
SRPK3	ENSG00000184343	1.823504	2.344521	1.34E-07
RPL3L	ENSG00000140986	1.600125	-2.55786	0.000279
NSUN5P2	ENSG00000106133	1.202103	2.178638	9.67E-07
BOP1	ENSG00000261236	1.019543	6.602543	2.33E-11
CELF5	ENSG00000161082	1.018047	-0.08126	0.022321
PRPF6	ENSG00000101161	1.010083	7.183575	6.40E-12
NSUN5	ENSG00000130305	0.949867	5.574691	9.77E-12
DDX10	ENSG00000178105	0.878494	4.496791	4.39E-09
BUD23	ENSG00000071462	0.826916	6.31088	9.31E-12
EIF3B	ENSG00000106263	0.776493	7.879597	2.07E-11

4.6 Gene signature of individual NMF cluster identified by LASSO logistic regression

To construct a gene panel and develop a classifier for individual NMF clusters, transcriptomic data of 14283 genes with high variance values (variance > 1) of 35 CRC organoid samples were used to train the least absolute shrinkage and selection operator (LASSO) logistic regression models. This technique effectively sets the coefficients of unimportant genes to zero and thus helps select a small number of genes that are important to the classification (Table 4). Subsequently, the LASSO model was evaluated on the test dataset of 19 samples. The overall accuracy is 78.95% with most of the classification errors occurred on samples from P2 (Figure 13).

Table 4 parameter values of model for signature genes of individual group

P1		P2		P3		P4	
ALPI	0.359582	ARHGEF17	0.12308	BRK1	0.006076	CGREF1	0.001625
CEBPA	0.001572	DBN1	0.016764	CADM1	0.091102	EVA1A	0.01896
CMAS	0.004479	EPHA2	0.009396	CD200	0.12079	FAM222A	0.06392
COL17A1	0.004827	FKRP	0.003578	EEF1A1P6	1.675654	HSPH1	0.027029
CTSS	0.033819	HDAC7	0.047626	RPL31	1.97E-06	NUP42	0.021811
FXYD3	0.000143	PPL	0.009162	SHISA6	0.029785	OSER1	0.045119
HMOX1	0.000932	PRSS22	0.007914	STAMBPL1	0.039384	POU5F1B	0.070034
PLOD2	0.037818	SHB	0.032714	TMEM258	0.00204		
SLC22A18A	0.00877	UBXN6	0.008412	TMEM51	0.006253		
SLC51A	0.003305						
TRIM7	0.043088						
WSCD1	0.021776						

```
Confusion Matrix and Statistics

          Reference
Prediction P1 P2 P3 P4
        P1  3  1  0  0
        P2  0  5  0  0
        P3  0  2  2  1
        P4  0  0  0  5

Overall Statistics

               Accuracy : 0.7895
                 95% CI : (0.5443, 0.9395)
    No Information Rate : 0.4211
    P-Value [Acc > NIR] : 0.001218

                  Kappa : 0.7175

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: P1 Class: P2 Class: P3 Class: P4
Sensitivity             1.0000    0.6250    1.0000    0.8333
Specificity             0.9375    1.0000    0.8235    1.0000
Pos Pred Value          0.7500    1.0000    0.4000    1.0000
Neg Pred Value          1.0000    0.7857    1.0000    0.9286
Prevalence              0.1579    0.4211    0.1053    0.3158
Detection Rate          0.1579    0.2632    0.1053    0.2632
Detection Prevalence    0.2105    0.2632    0.2632    0.2632
Balanced Accuracy       0.9688    0.8125    0.9118    0.9167
```

Figure 13 confusion matrix of prediction model

4.7 candidate gene markers selection for ribosome biogenesis

Since a unique characteristic of the P4 group, which is associated with resistance to chemotherapy treatment, is the up-regulation of ribosome biogenesis, another LASSO model was built to specifically classify the P4 group using expression profiles of 49 DEGs of P4 that are annotated with ribosome biogenesis pathway. This yielded a 7-gene signature for ribosome biogenesis (Table 5) that can classify the P4 group with 88.24% accuracy and 66.67% specificity on the test dataset.

Table **5** signature genes of ribosome biogenesis obtained from prediction model

Ribosome biogenesis related genes	Coefficient	log2FC
BOP1	0.0757	1.0195
GTPBP4	0.0295	0.6785
MPHOSPH10	0.2238	0.3234
PNO1	0.0334	0.7789
POP7	0.0629	0.3585
RRS1	0.1065	0.5812
TBL3	0.1007	0.2749

CHAPTER V

DISCUSSION

Genomic instability of CRC organoids in this study demonstrates two different genetic pathways in the development of sporadic CRC. Firstly, chromosomal instability pathway (47.37%) which is characterized by wide-spread alterations in chromosome number. Secondly, microsatellite instability pathway (10.53%) which exhibit defect in DNA mismatch repair system. In addition, somatic mutation profiles of these organoids exhibit mutations in several genes in adenoma-carcinoma sequence including APC gene as an early event, then activating KRAS mutation, and loss-of-function of TP53 tumor suppressor gene. Likewise, diverse mutation patterns of genes associated with CRC also identified in these organoids. However, the mutation frequencies in this study are lower frequency than the results found in TCGA database. This might be according to difference of disease stage and specimen between this study and TCGA database. However, this result exhibited genetically heterogeneous of these cancer organoids which consistent with that were found in colorectal cancer.

Application of NMF method to CRC organoids transcriptomic data able to be classified molecular subtypes into four groups with different molecular characteristics. P1: lipid metabolism, P2: WNT7A, B and high TGF-β pathway activation, P3: highly expressed TGF-β pathway, and P4: DNA repair upregulation, stem cell like subtype and ribosome biogenesis. These characters correlated with CMS2-4 while CMS1 are not enriched in any organoid group. This may cause from organoid sample are consist of epithelial cells without immune infiltrated cells thus including of MSI status might be necessary for additional marker. The classification results of this study demonstrated inconsistency among CMS subtypes prediction method in assigning CMS subtype from CRC organoid transcriptomic data. This result suggests that non-cancer transcripts from whole cancer tissues are required for current CMS classification method algorithm.

Molecular classification using transcriptomic data of CRC organoid which were particularly consisted of epithelial cells provide intrinsic molecular patterns of individual organoid grouping. Interestingly, P1 obviously presents lipid metabolism alteration correlated with CMS3. Moreover, this study able to divide TGF beta high associated with CMS4 into two group (P2 and P3) which were likely to metastasis by different mechanism. Besides, molecular features of P4 stem cell-like group such as high expression of WNT and MYC genes and chromosomal instability were

consistent to characteristics of CMS2, but previous CMS study demonstrated cancer stem cells (CSCs) traits were found in CMS4. This result might indicate sub-population of CMS2 which behave as CSCs of CRC. Furthermore, by using this approach stem cell like features, DNA repair and ribosome biogenesis pathway were identified in P4 radiotherapy resistance group.

Majority of enriched pathways of P1 involved in lipid and cholesterol metabolism which function in the cells by providing energy storage, structural component of cell membrane and messengers of metabolic signaling for the sake of cell proliferation (59). Furthermore, along the cancer transformation process, the acquisition of pro-survival abilities is an essential step that allow cancer cells to adapt to harsh cancer microenvironment and their contribution to cancer pathogenesis and progression (60, 61). Thus, cancer cells rewire their metabolism to acquire the fittest metabolic rate for homeostasis of cancer. Interestingly, it has been proposed that lipid metabolism alterations were found association with growth of primary tumor but also conducting tumor progression and metastasis (59). Thus, deeply investigate key enzymes involved in lipid metabolism can contribute to the development of the targeted therapy. However, sugar metabolism alterations together with KRAS mutation involving in metabolism rewiring of cancer were not found in this group. Thus, further investigation cause of lipid metabolism upregulation in this group is challenged.

The molecular characteristic of P2 and P3 are still ambiguous but upregulation of TGFβ pathway of this group were found. Hence, withdrawal of growth factors might be needed to further distinguish these groups.

For organoids in P4, most of it are present CIN phenotype involving the classic adenoma-carcinoma sequence characterized by a characteristic set of mutations in specific genes including APC, KRAS and TP53 (62, 63). Moreover, loss of APC function leads to hyperactivation of Wnt/beta-catenin signaling which regulate growth advantages in epithelial cells and it also considered as an early event in CRC tumorigenesis (64). This result support the characteristic of Wnt pathway hyperactivation of P4 organoids. Besides, the intestinal stem cells located in the base of the colon crypts are maintained in their undifferentiated state by wnt signaling pathway which contributes not only to the survival of normal stem cells but also to the survival of cancer stem cells. Activation of this pathway due to APC mutation leads to retention of a stem cell phenotype, which prevents them from migrating to top of the colon crypts to be discarded. This also agree with the intestinal stem cell-like properties of this group. Then, the aggregation of

undifferentiated cells in the colonic crypts eventually results in the formation of a polyp. Subsequently, accumulation of additional mutations in genes such as KRAS and TP53 may finally lead to carcinoma (65).

DNA damaged repair pathways play an important role in the maintenance of genome stability and integrity through correcting the impaired DNA that may contribute to carcinogenesis (56). Activation of the Fanconi anemia (FA) pathway occurs as a result of DNA replication of DNA damage, especially the damage triggered from DNA crosslinking agents. On the other hand, highly expression of FA genes and DNA damage repair capacity is helpful for relieving DNA lesions and chromosomal abnormalities that occur together with rapid proliferation of cancer cells. This was found pervasively in cancers and it was associated with chemoresistance (66). Moreover, it has been reported that tumor with high level of DNA damage repair related genes exhibited resistance to Cisplatin which is one of the most widely used chemotherapeutic drugs (67). Interestingly, the chemotherapy currently used in medical regiments for CRC patients including oxaliplatin, irinotecan, and 5-FU are directly or indirectly induced DNA damage which are recognized by specific DNA repair pathways. Thus, further testing of chemotherapy drugs in these CRC organoids are needed. Furthermore, these data indicate the need for development of gene markers to evaluate DNA repair capacity (DRC) of tumor cells which has been known to associated with chemo- and radiotherapy resistance (68).

For addition characteristic of P4, increased ribosome biogenesis in cancer cell is required to support cell growth and cell proliferation due to hyperproliferative cells were perturbed homeostasis of energy and increase protein synthesis activity. Moreover, continuous renewing of colon epithelium depends on self-renewal of stem cell, differentiation and proliferation activities which are maintain by the process of cell growth, division, protein synthesis and ribosome. Importantly, previous study found that in mouse colon organoids transcriptomic data, ribosome biogenesis signature was increased accompanied by differentiation of intestinal stem cells. Additionally, ribosome biogenesis factor and DNA-binding proteins are grouped together as nucleolus localized proteins. Majority of these proteins involved in DNA repair processes as well (69).

Intriguingly, ribosome biogenesis pathway associated with several mechanism of cancer cells. Thus, this pathway might be proposed as biomarker and candidate target for cancer treatment especially in radiotherapy resistance group. It has been studied that the clinical alkylating agent

Oxaliplatin does not induce cancer cell death through DNA damage but through ribosome biogenesis inhibition instead (70). These data support development of new drugs designed to selectively induce inhibition of ribosome biogenesis without the genotoxic effects which were currently used as anticancer drugs. Currently, ribosome biogenesis inhibitors were developed such as CX-3543 molecule disrupting the interaction of rDNA G-quadruplexes with a nucleolar protein necessary for Pol I transcription (71), CX-5461 inhibitor of rRNA transcription which presently under phase I clinical studies (72, 73), and BMH-21 molecule repressing RNA polymerase I transcription without causing DNA damage (74). These inhibitors have two main benefits. Firstly, they do not affect resting cells because the long half-life of cytoplasmic ribosomes and second, they can induce the apoptotic death of cancer cells, especially cancer with hyperactivation of ribosome biogenesis (58). Due to specific action mechanism these inhibitors can be combined with other anticancer drugs which act through different cytotoxic pathways such as energy related pathway to ensure a sufficient cancer cell destruction.

To detect ribosome biogenesis alteration in cancer cells, the silver staining of Nucleolar Organizer Regions (AgNOR) was used as a simple visualization. The AgNOR substitutes different argyrophilic nuclear proteins such as nucleolin and fibrillarin which are essential regulators of ribosome biogenesis (75). Towards AgNOR distribution was discovered to be associated with the nucleolus size which is linked to the cancer growth rate. Unfortunately, the AgNOR staining reaction is complicated to be automatized and requires laboratory technicians thus this approach is still not officially recommended in tumor pathology (58). This data reveals the need for the development of molecular markers to detect abnormality of ribosome biogenesis that could be applied in clinical practice. Accordingly, a seven-gene signature from supervised machine learning in this study might be developed as a prediction marker to classify samples with aberration in the ribosome biogenesis pathway. Likewise, candidate genes from co-expression analysis of this pathway provide additional results and have the potential as alternative gene markers as well. However, increase sample size and validation to other datasets are needed in further studies. Finally, to develop detection of ribosome marker using real-time quantitative polymerase chain reaction (RT- qPCR) method which easily use in clinical operation, additional PCR data collection of that gene signature is required for prediction model construction and indicate cutoff value.

In summary, unsupervised clustering of CRC organoid transcriptomic data reveals four organoid groups with different molecular characteristics. These data lead to pathway specific drug testing in each group of CRC organoids. Importantly, samples with stem cell-like properties were

clustered together in P4. This group highly expressed DNA repair pathways, hyperactivated ribosome biogenesis and performed radiation resistance. Thus, we purpose ribosome biogenesis pathway as a potentially alternative target for treatment and suggest gene signature of this pathway to predict organoids with radiation resistance.

REFERENCES

1. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nature Reviews Gastroenterology & Hepatology. 2019;16(12):713-32.

2. SEER Cancer Statistics Review [Internet]. 2019. Available from: https://seer.cancer.gov/csr/1975_2016/.

3. Sun D, Chen J, Liu L, Zhao G, Dong P, Wu B, et al. Establishment of a 12-gene expression signature to predict colon cancer prognosis. PeerJ. 2018;6:e4942-e.

4. Wang W, Kandimalla R, Huang H, Zhu L, Li Y, Gao F, et al. Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. Seminars in Cancer Biology. 2019;55:37-52.

5. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine. 2015;21(11):1350-6.

6. Dunne PD, Alderdice M, O'Reilly PG, Roddy AC, McCorry AMB, Richman S, et al. Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification. Nature Communications. 2017;8(1):15657.

7. Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. Genes & Diseases. 2019.

8. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Scientific Reports. 2017;7(1):16618.

9. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nature Communications. 2017;8(1):15107.

10. Lau HCH, Kranenburg O, Xiao H, Yu J. Organoid models of gastrointestinal cancers in basic and translational research. Nature Reviews Gastroenterology & Hepatology. 2020;17(4):203-22.

11. Sato T, Stange De Fau - Ferrante M, Ferrante M Fau - Vries RGJ, Vries Rg Fau - Van Es JH, Van Es Jh Fau - Van den Brink S, Van den Brink S Fau - Van Houdt WJ, et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. (1528-0012 (Electronic)).

12. Xie Y-H, Chen Y-X, Fang J-Y. Comprehensive review of targeted therapy for colorectal cancer. Signal Transduction and Targeted Therapy. 2020;5(1):22.

13. Nagtegaal ID, Quirke P, Schmoll H-J. Has the new TNM classification for colorectal cancer improved care? Nature Reviews Clinical Oncology. 2012;9(2):119-23.

14. Schlicker A, Beran G Fau - Chresta CM, Chresta Cm Fau - McWalter G, McWalter G Fau - Pritchard A, Pritchard A Fau - Weston S, Weston S Fau - Runswick S, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. (1755-8794 (Electronic)).

15. Marisa L, de Reyniès A Fau - Duval A, Duval A Fau - Selves J, Selves J Fau - Gaub MP, Gaub Mp Fau - Vescovo L, Vescovo L Fau - Etienne-Grimaldi M-C, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. (1549-1676 (Electronic)).

16. Sadanandam A, Lyssiotis Ca Fau - Homicsko K, Homicsko K Fau - Collisson EA, Collisson Ea Fau - Gibb WJ, Gibb Wj Fau - Wullschleger S, Wullschleger S Fau - Ostos LCG, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. (1546-170X (Electronic)).

17. De Sousa EMF, Vermeulen L Fau - Fessler E, Fessler E Fau - Medema JP, Medema JP. Cancer heterogeneity--a multifaceted view. (1469-3178 (Electronic)).

18. Budinska E, Popovici V Fau - Tejpar S, Tejpar S Fau - D'Ario G, D'Ario G Fau - Lapique N, Lapique N Fau - Sikora KO, Sikora Ko Fau - Di Narzo AF, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. (1096-9896 (Electronic)).

19. Roepman P, Schlicker A Fau - Tabernero J, Tabernero J Fau - Majewski I, Majewski I Fau - Tian S, Tian S Fau - Moreno V, Moreno V Fau - Snel MH, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. (1097-0215 (Electronic)).

20. Linnekamp JF, Hooff SRv, Prasetyanti PR, Kandimalla R, Buikhuisen JY, Fessler E,

et al. Consensus molecular subtypes of colorectal cancer are recapitulated in in vitro and in vivo models. Cell Death Differ. 2018;25(3):616-33.

21. Lenz H-J, Ou F-S, Venook AP, Hochster HS, Niedzwiecki D, Goldberg RM, et al. Impact of consensus molecular subtype on survival in patients with metastatic colorectal cancer: results from CALGB/SWOG 80405 (Alliance). Journal of Clinical Oncology. 2019;37(22):1876.

22. Menter DG, Davis JS, Broom BM, Overman MJ, Morris J, Kopetz S. Back to the colorectal cancer consensus molecular subtype future. Current gastroenterology reports. 2019;21(2):5.

23. Fontana E, Eason K, Cervantes A, Salazar R, Sadanandam A. Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. (1569-8041 (Electronic)).

24. Purcell RV, Schmeier S, Lau YC, Pearson JF, Frizelle FA. Molecular subtyping improves prognostication of Stage 2 colorectal cancer. BMC Cancer. 2019;19(1):1155.

25. Dunne PD, McArt DG, Bradley CA, O'Reilly PG, Barrett HL, Cummins R, et al. Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. Clinical Cancer Research. 2016;22(16):4095-104.

26. Alderdice MA-O, Richman SD, Gollins S, Stewart JP, Hurt C, Adams R, et al. Prospective patient stratification into robust cancer-cell intrinsic subtypes from colorectal cancer biopsies. (1096-9896 (Electronic)).

27. Allen WL, Dunne PD, McDade S, Scanlon E, Loughrey M, Coleman H, et al. Transcriptional subtyping and CD8 immunohistochemistry identifies poor prognosis stage II/III colorectal cancer patients who benefit from adjuvant chemotherapy. JCO Precis Oncol. 2018;2018:10.1200/PO.17.00241.

28. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503-11.

29. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

30. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix

factorization. Nature. 1999;401(6755):788-91.

31. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences. 2004;101(12):4164.

32. Fujii M, Shimokawa M, Date S, Takano A, Matano M, Nanki K, et al. A colorectal tumor organoid library demonstrates progressive loss of niche factor requirements during tumorigenesis. Cell stem cell. 2016;18(6):827-38.

33. Weeber F, Ooft SN, Dijkstra KK, Voest EE. Tumor Organoids as a Pre-clinical Cancer Model for Drug Discovery. Cell Chemical Biology. 2017;24(9):1092-100.

34. van de Wetering M, Francies HE, Francis JM, Bounova G, Iorio F, Pronk A, et al. Prospective derivation of a living organoid biobank of colorectal cancer patients. Cell. 2015;161(4):933-45.

35. Sato T, Stange DE, Ferrante M, Vries RG, Van Es JH, Van Den Brink S, et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology. 2011;141(5):1762-72.

36. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. 2017.

37. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.

39. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491-8.

40. Mayakonda A, Koeffler HP. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. BioRxiv. 2016:052662.

41. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Computational Biology. 2016;12(4):e1004873.

42. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics. 2014;30(7):1015-6.

43. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal; Vol 17, No 1: Next Generation Sequencing Data Analysis. 2011.

44. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34(5):525-7.

45. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. (1548-7105 (Electronic)).

46. Wickham H. ggplot2-Elegant Graphics for Data Analysis. Springer International Publishing. Cham, Switzerland. 2016.

47. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540-2.

48. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11(1):367.

49. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164-9.

50. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545.

51. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. (1362-4962 (Electronic)).

52. Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, et al. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis. 2019;8(9):44.

53. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1.

54. Kolde R. Pheatmap: Pretty heatmaps (version 1.0. 12). Google Scholar. 2019.

55. Kosinski C, Li Vs Fau - Chan ASY, Chan As Fau - Zhang J, Zhang J Fau - Ho C, Ho

C Fau - Tsui WY, Tsui Wy Fau - Chan TL, et al. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. (0027-8424 (Print)).

56. Li L-y, Guan Y-d, Chen X-s, Yang J-m, Cheng Y. DNA Repair Pathways in Cancer Therapy and Resistance. Frontiers in Pharmacology. 2021;11:2520.

57. Merlos-Suárez A, Barriga FM, Jung P, Iglesias M, Céspedes MV, Rossell D, et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. Cell stem cell. 2011;8(5):511-24.

58. Slimane SN, Marcel V, Fenouil T, Catez F, Saurin J-C, Bouvet P, et al. Ribosome biogenesis alterations in colorectal cancer. Cells. 2020;9(11):2361.

59. Fernández LP, Gómez de Cedrón M, Ramírez de Molina A. Alterations of Lipid Metabolism in Cancer: Implications in Prognosis and Treatment. Frontiers in Oncology. 2020;10(2144).

60. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.

61. Mouchiroud L, Eichner LJ, Shaw RJ, Auwerx J. Transcriptional coregulators: fine-tuning metabolism. Cell metabolism. 2014;20(1):26-40.

62. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5):759-67.

63. Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. Gastroenterology. 2010;138(6):2059-72.

64. Aghabozorgi AS, Bahreyni A, Soleimani A, Bahrami A, Khazaei M, Ferns GA, et al. Role of adenomatous polyposis coli (APC) gene mutations in the pathogenesis of colorectal cancer; current status and perspectives. Biochimie. 2019;157:64-71.

65. Armaghany T, Wilson JD, Chu Q, Mills G. Genetic alterations in colorectal cancer. Gastrointest Cancer Res. 2012;5(1):19-27.

66. Liu W, Palovcak A, Li F, Zafar A, Yuan F, Zhang Y. Fanconi anemia pathway as a prospective target for cancer intervention. Cell Biosci. 2020;10(1):39.

67. Oliver TG, Mercer KL, Sayles LC, Burke JR, Mendus D, Lovejoy KS, et al. Chronic cisplatin treatment promotes enhanced damage repair and tumor progression in a mouse model of lung cancer. Genes & development. 2010;24(8):837-52.

68. Vodicka P, Vodenkova S, Buchler T, Vodickova L. DNA repair capacity and response to treatment of colon cancer. Pharmacogenomics. 2019;20(17):1225-33.

69. Ogawa L, Baserga S. Crosstalk between the nucleolus and the DNA damage response. Molecular bioSystems. 2017;13(3):443-55.

70. Gaviraghi M, Vivori C, Tonon G. How Cancer Exploits Ribosomal RNA Biogenesis: A Journey beyond the Boundaries of rRNA Transcription. Cells. 2019;8(9):1098.

71. Rickards B, Flint S, Cole MD, LeRoy G. Nucleolin is required for RNA polymerase I transcription in vivo. Molecular and cellular biology. 2007;27(3):937-48.

72. Bywater MJ, Poortinga G, Sanij E, Hein N, Peck A, Cullinane C, et al. Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53. Cancer cell. 2012;22(1):51-65.

73. Drygin D, Lin A, Bliesath J, Ho CB, O'Brien SE, Proffitt C, et al. Targeting RNA polymerase I with an oral small molecule CX-5461 inhibits ribosomal RNA synthesis and solid tumor growth. Cancer Res. 2011;71(4):1418-30.

74. Peltonen K, Colis L, Liu H, Trivedi R, Moubarek MS, Moore HM, et al. A targeting modality for destruction of RNA polymerase I that possesses anticancer activity. Cancer cell. 2014;25(1):77-90.

75. Pelletier J, Thomas G, Volarević S. Ribosome biogenesis in cancer: new players and therapeutic avenues. Nature Reviews Cancer. 2018;18(1):51-63.

VITA

NAME	Pattarin Nuwongsri
DATE OF BIRTH	21 July 1993
PLACE OF BIRTH	Chonburi, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn university