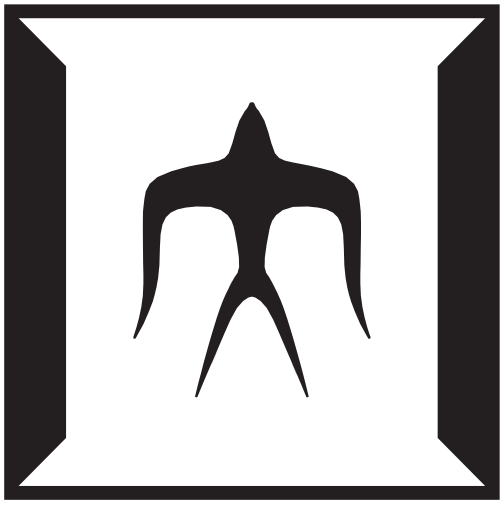
論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Metagenome and Metabolome Analysis of Japanese Lynch Syndrome Cohort
著者(和文)	FELIX SALIM
Author(English)	Felix Salim
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第12838号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:山田 拓司,伊藤 武彦,北尾 彰朗,本郷 裕一,平沢 敬
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第12838号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

博士論文

令和6年度

Metagenome and Metabolome Analysis of Japanese Lynch Syndrome Cohort



東京工業大学
Tokyo Institute of Technology

東京工業大学 生命理工学院

生命理工学系　生命理工学コース

Felix Salim

指導教員: 山田　拓司　准教授

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to express my gratitude to my supervisor Associate Professor Takuji Yamada for this research opportunity. He provided comfortable research environment and valuable inputs for the whole duration of my stay in this research team, encouraging independent research and methodology. Furthermore, he encouraged me to nurture scientific curiosity and open mind during my PhD course. I believe his guidance and my time spent in this laboratory would prove to be valuable for my future career.

Next, I would like to thank Dr. Sayaka Mizutani, Dr. Pande Putu Erawijantari and Dr. Hirotsugu Shiroma for their tutelage and advice in this project. Dr. Sayaka Mizutani provided valuable advice on the mechanism of CRC, research idea and methodology, as well as valuable inputs on scientific writing and presentations. Dr. Pande Putu Erawijantari provided mentoring on bioinformatics and statistical analysis, research idea, how to perform research works, as well as how to interpret and presents research results. Dr. Hirotsugu Shiroma performed metagenomic data preprocessing for the samples used in this research and helped me understand the details of metagenomic data preprocessing and data management.

Study subjects were collected in National Cancer Research Center Japan with the help of our collaborators. I would like to especially thank Dr. Shinichi Yachida and Dr. Satoshi Shiba for essential support on medical knowledge, sample collection, clinical data collection and for providing us with valuable data utilized in this research. Dr. Shinichi Yachida (Osaka University Graduate School of Medicine) provided valuable medical insights on mechanism of CRC and on research direction. Dr. Satoshi Shiba provided valuable support on data collection and in determining sample inclusion/exclusion criteria. I would also like to thank Dr. Takeshi Nakajima (National Cancer Center Hospital) for valuable insights to mechanism of Lynch Syndrome patients and how they are different to sporadic CRC patients.

I would like to thank Mr. Ryza Rynazal and Mr. Kota Fujisawa for valuable discussion about machine learning and the importance of explainable ML, Mr. Daisuke Suzuki for active discussion on the putative role of gut microbiota in CRC, Mrs. Yukiko Watanabe and Mrs. Nobuko Matsunami for her help and support in administrative work, Mr. Hirotaka Tsuchikura for technical support and server management, Mr. Takeshi Sasayama for management of subject's questionnaire data, Mrs. Etsuko Terajima for help in understanding metabolic pathways. I would like to thank everyone in Yamada Laboratory for their tutelage, discussions and supports on research, technical and personal problems.

I would like to thank MEXT for providing the scholarship enabling me to come to Japan to conduct my study and provide sufficient funding to let me focus on my PhD programme. Finally, I would like to thank my parents and sister for their patience, financial and mental support allowing me to focus on my study.

Chapter 1
Introduction
1.1 Background
1.1.1 Human microbiome and "omics" analysis

The human body is host to a huge variety of microorganism defined as the human microbiota, which have been shown to perform essential functions for the host development and health[1,2]. The human microbiota is constituted of bacteria, archaea, eukaryotes, and viruses, with the bacteria being one of the most extensively constituent of the human microbiota. The whole human microbiota, their genomes (i.e., genes), and the surrounding environmental conditions is defined as the human microbiome[3]. The human microbiome has been associated with many host conditions, such as age, sex, diet and diseases[4,5]. Associations between human microbiome and disease, such as colorectal cancer[6–8] and inflammatory bowel disease (IBD)[9–11], has been researched extensively, with recent studies focusing more on the causality of such associations[12–14]. Furthermore, there are many ongoing translational research for human microbiota-based therapy and diagnosis[15–19].

One of the most commonly utilized method to study the human microbiome is "omics" analysis, especially metagenomics. Metagenomics allows researchers to study the composition and potential function of the microbiota, including unculturable microbiota[20–22]. The standard protocol of metagenomics is 1) DNA extraction, 2) sample and library preparation, 3) DNA sequencing, 4) data processing, 5) data analysis[20,23]. Potential limitation of metagenomics is that signals from unexpressed DNA sequences, such as those that belong to dead members, may be picked up. Thus, other "omics" analysis, such as metatranscriptomics and metabolomics, is used in complement of metagenomics to get a better understanding of the studied environment[24].

In metabolomics, metabolites and small molecules within a biological sample was extracted and quantified[25,26]. The standard protocol of metabolomics is 1) extract analytes from the sample, 2) analytes separation (liquid/gas chromatography, capillary electrophoresis), 3) quantification (mass spectrometry, nuclear magnetic resonance), 4) data processing, 5) data analysis. The final metabolite profile depends on the combinations of sample extraction, separation and quantification protocols used; thus, it is important to customize the protocol for sample types and metabolites of interest. One limitation of metabolomics in microbiome analysis is that it is difficult to separate between host and microbiota derived metabolites.

"Omics" analysis allows researchers to study the microbiome in a culture-free way, bypassing conventional limitations of unculturable bacteria and low throughput experiments. In this study, metagenomics and metabolomics were performed on the fecal sample of Japanese Lynch Syndrome cohort, a hereditary colorectal cancer (CRC) cohort.

1.1.2 Human gut microbiota and colorectal cancer associations

Carcinogenesis model of CRC has been described as an adenoma-carcinoma sequence, where accumulations of gene mutation caused carcinogenesis progression[27–29]. The order of gene mutation is not deterministic, but some mutations were strongly associated with specific stages. For example, mutation in *APC* were associated with colorectal polyp formation, while *KRAS* mutation were associated with early adenoma transformation[28,29]. Most of such gene mutations trigger signalling cascade, causing uncontrolled cell proliferation and other hallmarks of cancer[28,29].

The human gut microbiome has been associated with CRC across all disease phases, including tumorigenesis, progression, and treatment response[6,13,30,31]. Specific microbes may produce harmful metabolites, inducing inflammations or gene mutations of colonic cell[32–34]. For example, enterotoxigenic *Bacteroides fragilis* produces enterotoxin which induced tumorigenesis[33], while *pks*$^{+}$ *Escherichia coli* derived colibactin induces genomic instability[34]. Others, like *Fusobacterium nucleatum* may invade colonic mucosa and induces inflammation by direct contact with the host cell via outer membrane protein[31,35]. Furthermore, gut microbiota derived metabolites such as butyrate are utilized as energy source by the colon cell, driving cell proliferation[36–38]. Under specific conditions, such cell proliferation have been linked to polyp formation[36,37,39,40].

The gut microbiota has also been linked with CRC progression[6,15,41–43]. For example, previous studies have shown distinct gut microbiota pattern associated with specific CRC disease stages, such as *Peptostreptococcus*, *Parvimonas* and *Fusobacterium* abundance trend to increase along with disease progression[6,42]. Furthermore, studies have shown *F. nucleatum* association with epithelial mesenchymal transition (EMT), potentially inducing cancer metastasis[44–49]. In vitro experiment shows further evidence of *F. nucleatum* colonized CRC tissue to move as individual cells rather than as a tissue[50].

Finally, recent studies have shown gut microbiota associations with treatment response[17,18,30,51]. For example, *F. nucleatum* have been shown to induce chemoresistance CRC via autophagy pathway activation[30]. Furthermore, *F. nucleatum* was also associated with immunotherapy response, albeit conflicting evidence on its exact roles[52–54].

Past studies have mostly focused on association studies, but recently the focus has shifted to mechanistic and translational studies. For mechanistic studies, causal relationship between human gut microbiomes and CRC have been shown in experiments with model mice, where inoculation of specific microbiota exacerbates or attenuates polyp formation and CRC progression[12,13]. However, microbiome studies on model mice may not translate well to human. Some studies tried to circumvent this with organoid experiments, utilizing human tissues rather than model mice[50,55–57].

Other challenges in elucidating the roles of human gut microbiome in CRC are due to the variation of CRC itself. Recent reports have shown that CRC differs between primary tumor sites and molecular subtypes[58,59]. Distinct patterns of human gut microbiome have been associated to differences in tumor site or molecular subtypes[60,61]. In similar fashion, host gene mutations have also been associated to human gut microbiome[4,62–65]. In the case of hereditary CRC, early loss-of-function of oncogenes or related genes may trigger intestinal environment change, affecting the human microbiome prior to CRC tumorigenesis.

1.1.3 Lynch Syndrome and colorectal cancer

As described in section 1.1.2, CRC progress as gene mutation accumulates[28,29]. CRC can be categorized as sporadic CRC and hereditary CRC based on how the gene mutation occurs[28,66]. In case of sporadic CRC, gene mutation occurs randomly on their somatic cell. However, gene mutation was inherited as an autosomal dominant germline mutation in hereditary CRC[28,66]. Hereditary CRC patients are at an increased risk of CRC due to the inherited mutation[28,66].

Lynch Syndrome (LS) is the most common cause of hereditary CRC, accounting for 3% of total CRC cases[66–69]. LS is caused by an autosomal dominant, heterozygous germline mutation in mismatch-repair (MMR) gene (*MLH1*, *MSH2*, *MSH6*, *PMS2*) or *EPCAM*[66–69]. LS mutations had different penetration rate according to mutated genes[70]. Other than CRC, LS also caused predisposition to other cancer, such as endometrial, stomach, small bowel, hepatobiliary system, upper urologic tract, ovary, glioblastomas, pancreas, breast, prostate and adrenocortical (LS-related cancer) [66–69]. Most LS study mainly focuses on colorectal and endometrial case, as those are the majority of LS-related cancer.

One of the hallmarks of LS is MMR deficiency (dMMR), caused by loss of function of the MMR mechanism[28,66–69]. LS subjects might be dMMR prior to any adenoma formation due to early mutation in the wild-type allele, leading to undetected dMMR colonic crypt[69,71]. MMR status can be diagnosed by immunohistochemistry of the MMR proteins or genetic test[69,72,73]. However, dMMR can also be seen in sporadic CRC, such as hypermethylated *MLH1* caused dMMR[74,75].

Related to dMMR, LS CRC patients are likely to have microsatellite instability[28,69,76,77]. Microsatellites are genomic regions consisting of repeated DNA sequences, which may be misaligned during DNA replication causing shorter or longer (instable) microsatellite. MMR proteins are responsible for repairing such error; thus, the mutations accumulate in dMMR subjects[28,69,76,77].

Other features of LS CRC are increased progress rate of adenoma-carcinoma sequences, better prognosis, chemoresistance, and better response to immune checkpoint inhibitors[28,69,78]. Most of these features have been associated to the dMMR colon cells in LS CRC. Due to dMMR, LS patients could not repair replication error inefficiently, resulting in an increased rate of gene

mutation accumulation, especially on regions with repeated sequences[28,69]. Increased gene mutation accumulation may accelerate the adenoma-carcinoma transformation process[28,69,71]. On the other hand, mutation in repeated sequence resulted in frame-shift peptides (FSP), which are novel antigen that can be targeted by the immune system, resulting in better prognosis[69,79–81]. Increased immunotherapy response rate may also be attributed to these novel antigens, as dMMR cancer cells became an easier target for the re-activated immune system compared to those with functioning MMR[82–85].

Recent studies study suggests that other than the conventional sporadic CRC pathway, LS CRC may progress through three other pathways (**Figure 1.1**)[69,71]. The first alternative is via sporadic adenomas that acquired secondary dMMR mutation. This is possible as although LS had a germline MMR mutation, loss-of-function only occurs after both alleles are mutated, in accordance to Knudson's double hit hypothesis. The second alternative is from dMMR colonic crypt that accumulated gene mutations and became carcinoma without *APC* mutation. In this second alternative, there is no adenoma step, resulting in hard-to-detect carcinoma. The third alternative is a derivative of the second pathway, but acquiring *APC* mutation midway, causing polyp formation. Other studies have reported that pathway preference is associated with mutated MMR gene.

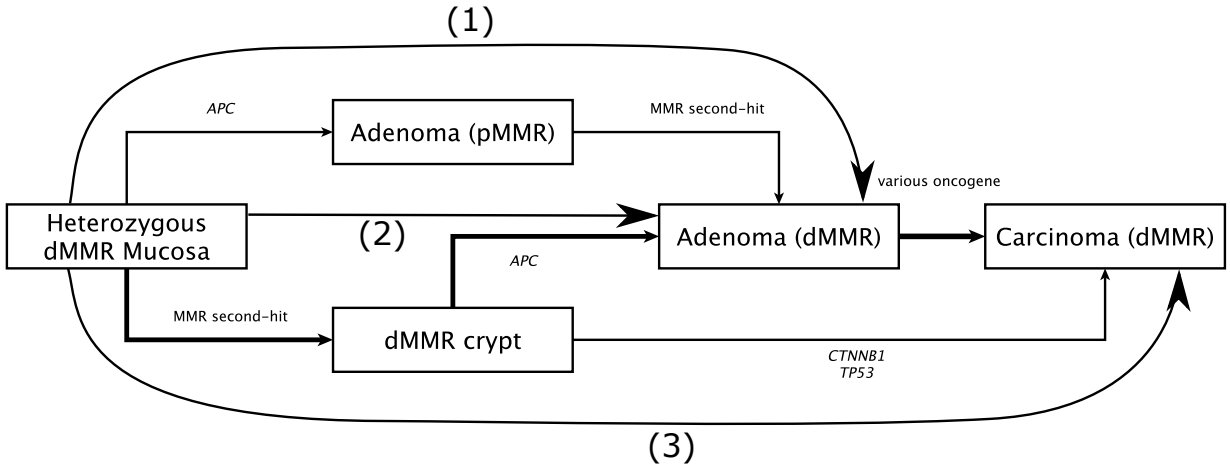


Figure 1.1 Putative CRC progression pathway in LS subjects

Three potential CRC progression pathways for LS subjects, with different point of MMR loss-of-function (LoF). 1) MMR LoF occurs after adenoma formation caused by stochastic mutation in other oncogenes, e.g., APC; 2) early MMR LoF due to stochastic event, followed by adenoma formation; 3) similar to the second pathway, with no apparent adenoma formation prior to carcinoma formation, possibly explaining undetected CRC cases

1.2 Motivation

Accumulating evidence showed associations between human gut microbiome and sporadic CRC[6–8]. However, there is limited knowledge on human gut microbiome association with hereditary CRC. Previous study has shown gut microbiota perturbations differ between dMMR CRC and non-

dMMR CRC, suggesting associations between MMR mutation with the human gut microbiota[86]. Thus, it is likely that LS CRC, characterized by MMR mutation, also has distinct human gut microbiota association.

Recent studies showed that human gut microbiome is also associated with CRC subject's treatment response[30,52,54]. Specifically, *Fusobacterium nucleatum* has been associated with chemoresistance and immune response therapy, albeit conflicting evidence for the latter[30,52,54]. Notably, *F. nucleatum* is one of the microbes associated with MMR mutation in the host, with reports of high *F. nucleatum* presence in MSI-high CRC patients[87,88] (**Figure 1.2**).

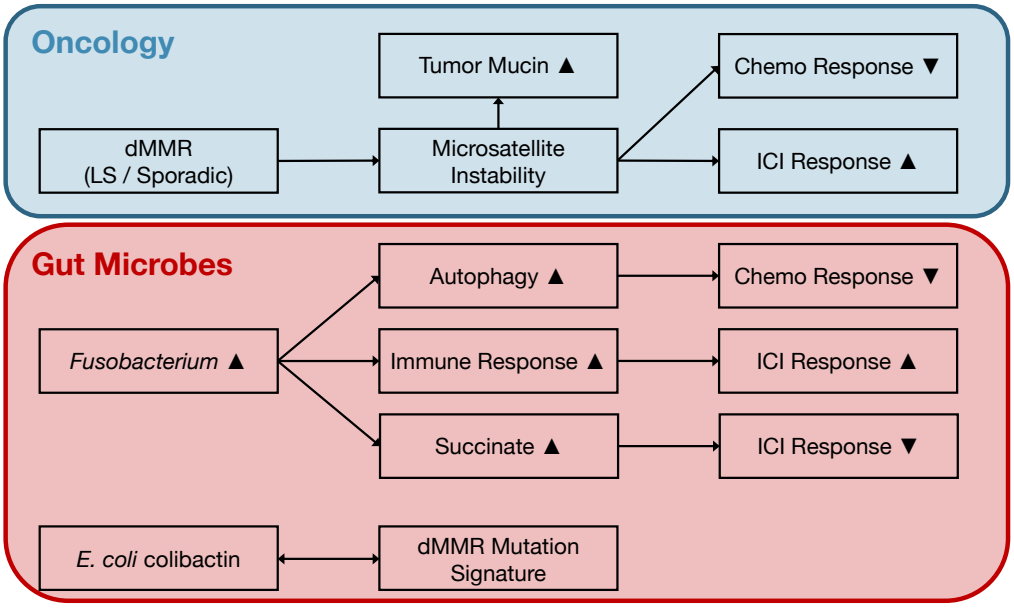


Figure 1.2 Known dMMR CRC characteristics and gut microbiota-CRC associations

From cancer histopathology point-of-view, dMMR/MSI-H CRC have been reported to be mucinous and have distinct response to treatment, regardless of dMMR/MSI-H origin (LS or sporadic MLH1 hypermethylation). From gut microbiota studies, *F. nucleatum* have been associated with MSI-H CRC, chemoresistance via autophagy activation, and ICI therapy response. However, the exact association of *F. nucleatum* with ICI response is still up-to-debate. Other than *F. nucleatum*, a study reports the mutational signature of colibactin, a genotoxin derived by specific *E. coli* strains, co-occurs with dMMR mutational signature.

Studies on mice intestinal microbiota suggested gut microbiota drives polyp formation in MSH2/APC mutant mice, likely by driving intestinal cell proliferation via butyrate production[39]. Furthermore, gut microbiota may be related to intestinal inflammation, which is also associated with CRC tumorigenesis[89–91].

Recently, distinct human microbiota structure of LS subjects has been reported[92–94], with one study reporting history of cancer as a confounding factor[94]. Additionally, inflammation related genes were reported to be weakly predictive for adenoma formation in LS subjects[95].

Nevertheless, previous studies have various limitations, such as low sample number, lack of non-LS comparisons, and lack of LS subjects with no prior tumor formation. Furthermore, most studies were done with 16S rRNA sequencing, thus there is limited information on the gut microbiota functions in LS subjects.

In this study, metagenome and targeted metabolome was performed on fecal samples collected from a Japanese LS cohort. The cohort consisted of LS subjects with no prior adenoma and carcinoma formation (LS controls), LS subjects with colonic adenoma formation (LS adenoma), LS subjects with colonic carcinoma formation (LS CRC), and LS subjects with colectomy records (LS surgery). In addition, a Japanese sporadic CRC cohort with identical sampling and sequencing protocol was utilized as a comparison cohort to identify potential LS-specific associations[6]. Furthermore, questionnaire data on subject's lifestyle and dietary habit was available for most subjects. Overall, the study aims to identify potential associations between human gut microbiome and CRC tumorigenesis or progression in LS subjects.

1.3 Contribution

This study provides novel insights into the microbiota composition and functions of LS subjects, particularly in relation with CRC progression. The study's results may serve as a framework to generate hypothesis on how host genetics affect gut microbiota and cascade into disease development.

As mentioned in section 1.1.3, LS is characterized by accelerated adenoma-carcinoma progression, better prognosis and better response to ICI therapy, which can be utilized as model case for dMMR CRC progression and treatment. Thus, findings of this study may provide clues on how gut microbiota and its function is linked to CRC progression and treatment response.

1.4 Dissertation Structure

The content is structured as follows. In Chapter 2, the cohort and methods used in this study was introduced, with an emphasis on the computational processing and analysis. In Chapter 3, I describe and discuss the study findings. Finally, the general summary of study findings and potential development for future research was introduced in Chapter 4.

1.5 Publications

1. **Salim, F.**, Mizutani, S., Shiba, S., Takamaru, H., Yamada, M., Nakajima, T., Yachida, T., Soga, T., Saito, Y., Fukuda, S., et al. (2024). *Fusobacterium* species are distinctly associated with patients with Lynch syndrome colorectal cancer. iScience 27, 110181. https://doi.org/10.1016/j.isci.2024.110181.

Chapter 2
Material and Methods

As mentioned in Chapter 1, whole-shotgun sequencing metagenomics and targeted metabolomics was performed on fecal samples collected from a Japanese LS cohort. In this chapter, details of the LS cohort utilized in this study, experimental protocols and computational methodologies is described. This project is a collaborative project between Tokyo Institute of Technology, National Cancer Center Japan, Osaka University, and Keio University. Subjects were enrolled by National Cancer Center Japan, followed by fecal sample and clinical data collections, including colonoscopy and questionnaire data. DNA sequencing was performed by sequencing service provider, while targeted metabolomics was performed in Keio University. DNA reads processing and annotation; downstream statistical analysis of metagenome and metabolite profiles was done in Tokyo Institute of Technology by the author. Results interpretation was carried out by intensive discussions with collaborators (professor and clinicians) from collaborating institutes (Osaka University, National Cancer Center Japan, Keio University and Tokyo Institute of Technology).

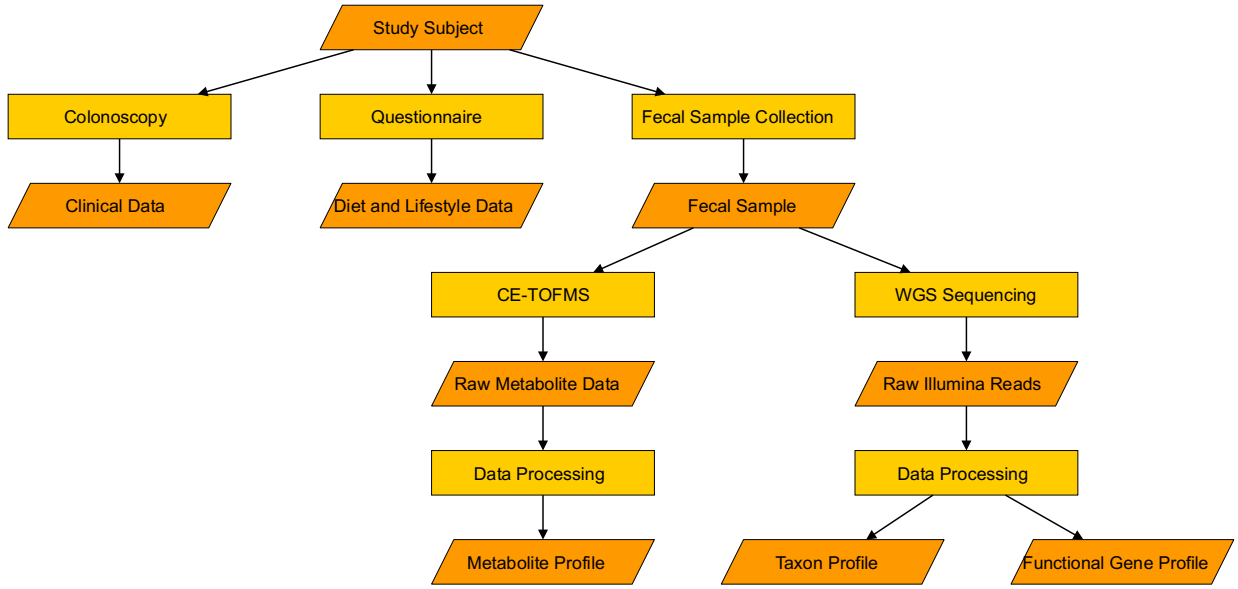


Figure 2.1 Data collection and processing overview

Study subjects were recruited in NCC Japan from incoming colonoscopy patients. Fecal sample, clinical data, and lifestyle questionnaire was collected by NCC collaborators. Fecal sample sequencing was performed by external service with Illumina platforms and the sequenced raw reads was processed and analyzed by author. Fraction of fecal sample was sent to Keio University collaborators for metabolite measurements.

2.1 Cohort overview

Patients undergoing colonoscopy at the National Cancer Center Hospital, Tokyo, Japan, were enrolled in the study, totaling 71 patients with Lynch Syndrome. The inclusion criterion was a positive diagnosis of a pathogenic mutation in germline MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*,

or *EPCAM* + *MSH2*). Patients with a history of gastrectomy were excluded. Patients' stool samples were collected after bowel preparation on the day of colonoscopy and stored at -80°C before further processing.

Patients with LS were grouped into adenoma (LS-adenoma), CRC (LS-CRC), and post-colectomy (LS-surgery) groups based on the colonoscopy results and clinical records. Those without any adenoma, carcinoma, or history of colectomy were defined as controls (LS control). The lifestyle and dietary data of some patients were acquired using questionnaires based on a previous study[96]. Potential confounding factors such as age, sex, mutated MMR genes, smoking habits, meat consumption, and dietary fiber consumption were determined from clinical and questionnaire data and compared across the groups (**Table 2.1**).

Table 2.1 Lynch Syndrome subject's clinical information

	LS-CTR[1]	LS-ADE[2]	LS-CRC[3]	LS-SUR[4]	*P**
Subject Number	18	23	11	19	
Age (mean (SD))	39.56 (11.13)	50.74 (13.70)	46.82 (19.11)	52.95 (13.33)	0.025
Sex = Male (%)	5 (27.8)	11 (47.8)	7 (63.6)	7 (36.8)	0.248
BMI[5] (mean (SD))	21.16 (2.47)	24.12 (4.89)	21.88 (3.94)	24.06 (4.57)	0.089
Brinkman Index (mean (SD))	97.65 (163.08)	161.33 (227.30)	344.00 (372.71)	235.79 (389.48)	0.178
Total dietary fiber intake (mean (SD))	11.68 (6.11)	10.39 (7.38)	14.66 (6.78)	11.28 (4.68)	0.406
Meat intake (mean (SD))	90.44 (48.59)	90.23 (87.41)	102.45 (74.44)	72.28 (44.34)	0.68
Mutated MMR[6] variants (%)					0.319
MLH1	6 (33.3)	10 (43.5)	2 (18.2)	7 (36.8)	
MSH2	11 (61.1)	10 (43.5)	5 (45.5)	7 (36.8)	
MSH2+EPCAM	0 (0.0)	1 (4.3)	1 (9.1)	1 (5.3)	
MSH6	0 (0.0)	2 (8.7)	0 (0.0)	2 (10.5)	
PMS2	1 (5.6)	0 (0.0)	3 (27.3)	2 (10.5)	
Colectomy history = yes (%)	0 (0.0)	8 (34.8)	3 (27.3)	19 (100.0)	<0.001
Colectomy part (%)					<0.001
none	18 (100.0)	15 (65.2)	8 (72.7)	0 (0.0)	
right	0 (0.0)	3 (13.0)	2 (18.2)	11 (57.9)	
left	0 (0.0)	2 (8.7)	1 (9.1)	4 (21.1)	
both	0 (0.0)	3 (13.0)	0 (0.0)	4 (21.1)	

[1]LS-CTR = Lynch Syndrome (LS) subjects without adenoma, carcinoma, or history of colectomy

[2]LS-ADE = LS subjects with colorectal adenoma
[3]LS-CRC = LS subjects with CRC
[4]LS-SUR = LS subjects with colectomy history, but no current colorectal adenoma or CRC
[5]BMI = body mass index
[6]MMR = mismatch repair
[*]One-way analysis of variance (one-way ANOVA) test was used for numerical variables and chi-square test was used for categorical variables

Data on patients with sporadic CRC were obtained from previously published studies with some modifications, specifically only non-LS healthy controls with no prior history of colorectal adenoma formation were included in this study[6]. Based on colonoscopy results, patients with sporadic CRC were grouped into five groups: (1) controls (no remarkable colonoscopic findings or records of adenoma); (2) multiple polypoid adenomas with low-grade dysplasia (more than three adenomas, mostly more than five adenomas); (3) intramucosal carcinoma (polypoid adenoma(s) with high-grade dysplasia), stage 0/pTis CRC (S0); (4) early CRC (Stages I or II); (5) advanced CRC (Stages III or IV). As previously mentioned, potential confounding factors such as age, sex, smoking habit, meat consumption, and dietary fiber consumption were determined from clinical and questionnaire data and compared across subject groups (**Table 2.2**).

Table 2.2 Sporadic CRC subject's demographic

	SP CTR	SP ADE	CRC S0	CRC S12	CRC S34	P^*
Subject number	112	67	73	110	75	
Age (mean (SD))	55.89 (13.24)	63.15 (9.12)	64.82 (7.54)	63.81 (9.34)	59.08 (10.92)	<0.001
Sex = male (%)	59 (52.7)	48 (71.6)	45 (61.6)	72 (65.5)	44 (58.7)	0.109
BMI (mean (SD))	22.43 (2.99)	23.09 (4.70)	23.33 (3.57)	22.82 (3.18)	23.13 (3.19)	0.451
Brinkman Index (mean (SD))	204.88 (379.10)	502.70 (428.94)	452.27 (513.02)	415.75 (516.19)	376.27 (469.59)	<0.001
Total dietary fiber intake (mean (SD))	14.20 (13.75)	12.66 (15.51)	13.35 (12.13)	12.33 (7.82)	11.76 (5.64)	0.652
Meat intake (mean (SD))	88.12 (102.00)	80.21 (57.93)	76.24 (58.98)	65.93 (46.42)	77.05 (112.35)	0.382

[*]One-way analysis of variance (one-way ANOVA) test was used for numerical variables and chi-square test was used for categorical variables

2.2 Subject's details and fecal sample collections

Study participants were part of a joint project with National Cancer Research Center Japan (NCC), Tokyo Institute of Technology (TITech), and Keio University. Fecal samples, clinical information and questionnaire data were obtained under informed consent of study participants and approved by each institute institutional review boards (NCC 2013-244, TITech 2014018).

Subjects with a positive gene test result (germline pathogenic or likely pathogenic variant of any MMR genes or EPCAM gene) was treated as LS subjects. Subject with gastrectomy surgery history or other hereditary CRC, e.g., familial adenomatous polyps (FAP), were excluded from this study.

Subjects were categorized by NCC experts based on colonoscopy results and previous clinical records. Subjects with colorectal carcinoma was assigned TNM classifications based on The Union for International Cancer Control (UICC) The TNM Classification of Malignant Tumours 8th edition. To simplify CRC classification and account for low sample number, equivalent AJCC stages were assigned and aggregated into three groups, 1) intramucosal CRC (Stage 0), 2) early CRC (Stage 1 and Stage 2), and 3) advanced CRC (Stage 3 and Stage 4).

Subjects were asked to fill questionnaires about lifestyle data, including dietary habits, medical records and family information (475 questions, 25 pages). Questionnaire were designed based on previous large Japanese prospective cohort study[96]. Potential confounding factors, such as fiber and meat intake or smoking habit, was extracted from the full questionnaire data. Smoking habit was quantified into Brinkman Index, an indicator of tobacco intake calculated from years of smoking and average cigarette packs consumption per year (obtained from questionnaire or medical records).

Subject's information was summarized into a table via R tableone package and manually modified for ease of view. One-way analysis of variance and chi-square test was performed on continuous variable and discrete variable, respectively.

Fecal samples were collected at first defecation after oral administration of bowel cleansing agents (one of MAGCOROL P, MOVIPREP or MUBEN). Sample collection was performed at the hospital on the day of colonoscopy. Fecal samples were placed directly on dry ice and stored at -80°C until further experiments.

2.3 Targeted metabolome quantification

Targeted metabolome quantification was performed by a collaborator from Institute for Advanced Biosciences, Keio University, based on a previously described protocol[97,98]. Fecal metabolites were extracted from ten milligrams of fresh thawed freeze-dried fecal samples suspended in 400 μL of 50% methanol in Milli-Q containing internal standards (20 μM each of methionine sulfone and D-camphor-10- sulfonic acid (CSA). The 3-mm zirconia beads (BioSpec Products, Bartlesville, OK, USA) and 100 mg of 0.1-mm zirconia/silica beads (BioSpec Products, Bartlesville, OK, USA) were added into the mixture then subjected to vigorous shaking using Micro Smash (TOMY, Nerima, Tokyo, Japan). The suspensions were centrifuged for 10 minutes at 1500 rpm, 10 times. The supernatant was transferred to a 5-kDa- cutoff filter column (Ultrafree MC-PLHCC 250/pk) for Metabolome Analysis (Human Metabolome Technologies, Tsuruoka, Yamagata, Japan). The

flow-through was dried under vacuum and the residue then was dissolved in 40 μL of Milli-Q water containing reference compounds (200 μM each of 3-aminopyrrolidine and trimesate). The levels of extracted metabolites were measured in both positive and negative modes by Capillary Electrophoresis Time-of-Flight Mass Spectrometry (CE-TOFMS). The CE-TOFMS experiments were carried out using an Agilent CE Capillary Electrophoresis System (Agilent Technologies, Santa Clara, CA, USA).

The raw data were processed for metabolite quantification using automatic integration software MasterHands (ver. 2.16.0.15). The annotations tables were generated based on the measurement of standard compounds and aligned with the datasets according to the similar m/z values and normalized migration time. The peak areas were then normalized against those of the internal standards methionine sulfone and CSA for cationic and anionic metabolites, respectively. Concentrations of each metabolite were calculated based on their relative peak areas and the concentrations of the standard compounds. For the analysis, concentrations below the detection limit were substituted with zero, and metabolites for which levels were below the detection limit in all of the samples were excluded. The metabolite concentration (nmol) was normalized using the fecal weight to obtain the amount of metabolite in each gram of sample (nmol/g). The profiles of metabolites were then processed for statistical analysis in Tokyo Institutes of Technology (by the author).

2.4 DNA extraction and whole genome shotgun sequencing

DNA were extracted from frozen fecal sample by bead beating method via GNOME® DNA Isolation Kit (MP Biomedicals). Quality of extracted DNA are assessed by measuring DNA spectrometry via 4200 TapeStation (Agilent Technology). Samples with low DNA concentration or accuracy are not used in this study. After final precipitation, DNA samples were resuspended in TE buffer and stored at -80°C before downstream procedures. All procedures were performed in National Cancer Center Japan.

Sequencing libraries were generated via Nextera XT DNA Sample Prep Kit (Illumina) from defrosted DNA sample. Library quality was confirmed via 4200 TapeStation (Agilent). Whole-genome shotgun sequencing was performed on the HiSeq2500 platform (Illumina), with 150-bp pair-ended sequencing. Targeted sequencing depth of WGS procedure is 5.0 Gb. Whole-genome shotgun sequencing procedures was carried out by a sequencing company.

2.5 Reads pre-processing

Raw reads from whole-genome shotgun sequencing may contain contaminated reads, sequencing error, and other source of bias. Furthermore, host genomes were not removed prior to sequencing. Therefore, raw reads were pre-processed prior to downstream assembly and annotation.

The preprocessing pipeline and related codes was constructed by previous lab members and have been reported in prior publication[6,99]. Briefly, reads with ambiguous bases, short length (length < 50 base pair), or low quality (average Phred Q < 25) was removed. Furthermore, adapters, primers, spike-in control (*phiX* DNA), and human genome was also removed. The full pre-processing pipeline is shown in **Figure 2.2**.

In detail, reads with ambiguous bases, shown as N in the FASTQ file, were removed with in-house Python script. Then, reads were mapped to *phiX* reference genome using Bowtie2 (v2.2.9)[100] with preset options' "-fast-local", keeping only unmapped reads. Next, adapter sequence and low-quality 3' terminal sequences were trimmed with cutadapt (v1.9.1)[101], with the following parameters: -a CTGTCTCTTATACACATCTCCGAGCCCACGAGAC -O 33 -q 17 for the forward primer sequence; -a CTGTCTCTTATACACATCTGACGCTGCCGACGA -O 32 -q 17 for the reverse primer sequence. Next, reads with length less than 50 base pair and reads with average quality value lower than 25 were discarded with in-house Python script. Human genomes (GRCh38, all chromosomes + mitochondria) were discarded in the same manner as *phiX* genome. Next, forward primer reads and reverse primer reads were matched by reads ID to remove unpaired reads with in-house Python script. Finally, pair-end FASTQ reads were concatenated and transformed into FASTA format with an in-house Python script.

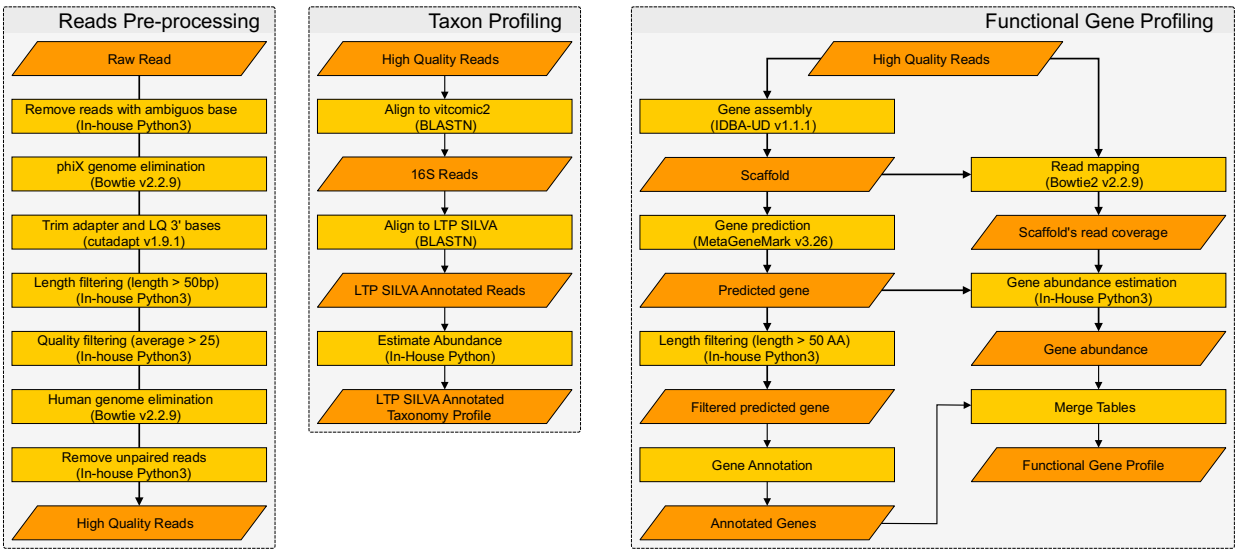


Figure 2.2 Data processing, taxon and functional gene profiling pipeline

Raw reads were pre-processed to remove sequencing artifacts and human genomes with a previously published pipeline. Pre-processed reads were then annotated with taxon information and functional gene information. Furthermore, taxon and functional gene abundance were estimated based on reads count and gene coverage.

2.6 Taxonomic profiling

As mentioned in section 1.1.1, metagenome sequencing is commonly used to study the microbiota composition of specific environments. In order to estimate the microbiota composition, sequenced

reads were annotated with taxon information up to species level (**Figure 2.2**). Furthermore, taxon abundance was estimated from the annotated reads count. In this study, 16S ribosomal RNA (rRNA)-based taxon annotations were performed. Specifically, SILVA LTP reference database was chosen as it is one of the more comprehensive 16S reference database and it was used in the sporadic CRC publication[102].

2.6.1 16S rRNA extraction

The ribosome, which is responsible for the protein synthesis in all living organisms, consisted of a large subset and a small subset. In prokaryotes, the large subset is called 23S and the small subset is called 16S, both of which have conserved and variable regions in its sequences. By aligning the sequenced 16S rRNA reads against a reference database of 16S rRNA variable regions, it is possible to estimate which prokaryote the reads belong to. Usually, a universal primer specific to variable regions of 16S rRNA were used to amplify the specific regions prior to sequencing. In this study, as whole genome shotgun is used, 16S rRNA were extracted by aligning high-quality reads against a full-length 16S rRNA reference database (VITCOMIC2[103]) with BLAST (version 2.2.30)[104] (cut-off: E-value $< 1 \times 10^{-8}$). Reads aligned with the database were utilized in downstream process as 16S rRNA reads.

2.6.2 SILVA LTP annotation

The extracted 16S rRNA reads were aligned to the 16S rRNA sequence database provided by The All-Species Living Tree (LTP) project of the SILVA database (version 123)[102] using BLASTn (cut-offs: E-value $< 1 \times 10^{-8}$, sequence identity $> 97\%$, alignment coverage $> 80\%$, bit score > 70)[104]. Alignment(s) with the highest sequence identity and bit score were used as aligned reads to estimate downstream abundance. In case of multiple alignments, the aligned reads were divided by the number of aligned taxa so that they could be "shared." Species-level relative abundance was computed per sample and was defined as the number of reads assigned to the species divided by the total number of LTP-aligned reads in the sample. Relative abundances at the genus level and higher taxonomic ranks were calculated as the sum of the relative abundances of all member species. The generated profiles are referred to as "species profile" and "genus profile" hereafter.

2.7 Functional profiling

2.7.1 Genome assembly, gene prediction and abundance estimation

Prior to gene annotation, high-quality reads were assembled into longer contigs and scaffolds with IDBA-UD[105]. Next, genes position in the scaffold were predicted with MetaGeneMark[106]. To remove potential false positives, genes were filtered by length (length > 50 amino acids) before downstream annotation (**Figure 2.2, 2.3**).

In order to consider gene length differences, gene's read coverage was used as gene abundance in downstream analyses. Read coverage was estimated by mapping high-quality reads to the

assembled scaffolds, then dividing the mapped read length by the gene length for each gene. Read mapping was done with Bowtie2 (v2.2.9)[100].

2.7.2 KEGG annotation

Length-filtered gene sequences were aligned against a reference database of prokaryotic complete genomes recorded in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[107] (obtained on 2023/03/31) using DIAMOND (version 2.1.2)[108] (cut-offs: sequence identity > 40, bit score > 70, coverage > 80). Alignment results were merged with estimated gene abundance table to calculate annotated gene abundance. For cases where query gene aligned to a single reference gene, read coverage was used as is. If the query aligned to multiple reference genes, the read coverage was "shared" over all corresponding genes by dividing the read coverage with the number of aligned genes. Finally, gene-level relative abundances were calculated per sample by dividing the read coverage of the gene by the sum of all gene read coverages. Relative gene abundance was further aggregated to the KEGG Orthology (KO) level. The generated profile contained 8,198 KO and referred to as the "KO gene profiles".

2.7.3 Fusobacterium nucleatum virulence factor annotation

Previous studies have reported associations between human gut microbiota and CRC, with *Fusobacterium nucleatum* (Fn) as one of the most well-studied CRC-associated gut microbes. In addition, previous studies have shown Fn was involved in CRC progression and treatment response, with several Fn genes identified as potential virulent factors. Higher Fn load is associated with MSI-H, dMMR and common molecular subtype (CMS) 1, all of which related to LS subjects. Therefore, I explored the metagenome for potential Fn virulent factors (Fn VF) to see if Fn VFs is enriched in LS subjects.

Reference amino acid sequences of Fn VFs (*fap2*, *fadA*, *fadA2*, *fadA3*, *FnDps*, *cbpF*, and *radD*) were obtained from a previous study[109]. To expand the search space for these virulence genes, reference genomes were used as queries for eggnog-mapper (version 2.1.9)[110] to obtain eggNOG (v5)[111] IDs, each representing an orthologous gene cluster. Obtained eggNOG IDs were then used as references for Fn VFs annotation. Next, annotation results were merged with the gene abundance table to calculate eggNOG abundance. If query gene was annotated with multiple eggNOGs, the read coverage was split evenly for all corresponding eggNOGs, identical with previously described KEGG annotation. Each eggNOG read coverage was then divided by the total sum of the eggNOG read coverages to obtain the relative abundance of each eggNOG. Finally, Fn VF-matched eggNOG IDs were extracted and utilized as Fn VF abundance in downstream analysis.

2.7.4 Colibactin biosynthesis cluster annotation

Previous studies have reported potential host DNA damage caused by gut microbiota, one of which is by *Eschericia coli* derived colibactin[34,112]. Colibactin have been shown to cause double strand

DNA damage and exacerbate inflammation associated CRC tumorigenesis[14,34]. Furthermore, a recent study has shown association between colibactin and dMMR related mutation signature, suggesting colibactin exacerbate dMMR related DNA mutation[112]. As dMMR is one of the hallmarks of LS[28,69], I explored the metagenome for colibactin biosynthesis cluster genes to see if colibactin production is higher in LS subjects.

UniProt IDs for the colibactin biosynthesis cluster genes were obtained from a previous study[113]. Next, length-filtered genes were aligned against the UniProt TrEMBL database (release 2022_05)[114] using DIAMOND (version 2.1.2)[108] (cut-offs: sequence identity > 40, bit score > 70, and coverage > 80). Alignment results were merged with gene abundance table to calculate UniProt gene abundance. Finally, reference-matched UniProt IDs were used for colibactin-producing gene abundance in the downstream analysis.

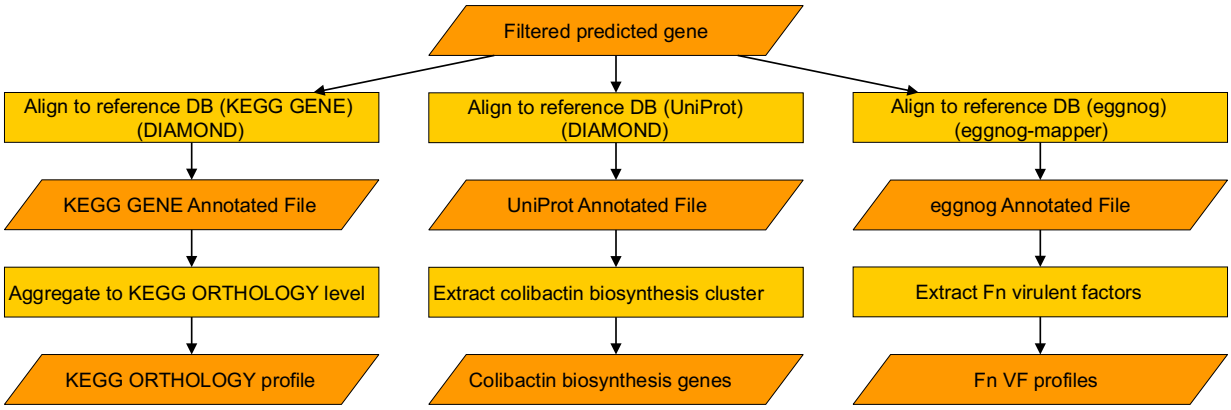


Figure 2.3 Functional gene annotation details

Genes predicted by **Figure 2.2** was annotated with KEGG Orthology information for general metabolism related functions. As KEGG was not comprehensive for putative virulence factor or specific metabolism pathway, UniProt and eggNOG was utilized for colibactin biosynthesis genes and *Fusobacterium nucleatum* virulence factor, respectively.

2.7.5 Secretion related protein annotation

Whether or not a protein or metabolites is secreted or not plays a major role in host-microbe and microbe-microbe interaction[115]. Similarly, outer membrane proteins are hypothesized to be involved in host-microbe and microbe-microbe interaction. One characteristic of secretion-related and outer membrane proteins is a short N-terminal amino sequence called signal peptides[116]. Several tools have been developed to predict secretion related protein based on signal peptides presence, one of which is signalP, which utilized a language model for secretion related protein prediction[117].

Length-filtered gene sequences were used as queries to signalP (v6.0)[117]. Predicted results were filtered for genes belonging to one of the five signal peptides category. Then, results were merged with KO abundance table to obtain KO-annotated secretion-related protein abundance profile.

2.7.6 Carbohydrate active enzyme (CAZy) annotation

One major role of the gut microbiota is to digest dietary fibers and produce short-chain fatty acids, which acts as energy source for colon epithelial cells[118,119]. Other than dietary fiber, gut microbiota was also known to degrade the intestinal mucus layer, which were composed by a variety of host glycans, such as mucin[118,120]. Recent studies showed the homeostasis of fiber degrading and mucin degrading gut microbiota may be associated with the host health[121].

To explore the fiber degrading and mucin degrading capacity of the gut microbiota in LS cohort, carbohydrate active enzymes (CAZy) annotation was done[122]. CAZy consisted enzymes which degrades or synthesis carbohydrates, which was categorized into 4 categories: 1) glycoside hydrolase (GH), 2) glycosyltransferase (GT), 3) polysaccharide lyase (PL), and 4) carbohydrate esterase (CE). Additionally, carbohydrate-binding module (CBM) was also included.

Length-filtered genes were used as queries for run_dbcan (v3.0.6), a command-line tool to annotate genes with dbCAN2[123]. Within run_dbcan, three annotation methods were used and annotation results that were consistent in at least 2 methods was used as the CAZy annotation. The annotation methods were: 1) query alignment against reference database with DIAMOND; 2) hidden Markov model (HMM) profile-based search with HMMER; 3) k-mer based search with eCAMI. In case conflicting annotation results were obtained, HMMER and DIAMOND consensus takes precedence over eCAMI results.

2.8 Downstream data analysis

2.8.1 Microbiota composition and structure analysis

The human gut microbiota composition has been associated with host health. There are several ways to analyze the human gut microbiota composition, one of the most common one is to quantify within-sample diversity (alpha diversity) and between-sample diversity (beta diversity)[124].

There are several methods used to quantify alpha diversity, one of which is the Shannon-Wiener index, which considers not only the number of organisms within a sample, but also considers the organism's abundance distribution. The Shannon–Wiener index was calculated on the LTP annotated taxonomic profiles at the species level ("diversity" function, index = "shannon," R vegan). The calculated index was compared between the LS and sporadic cohorts using the Mann–Whitney U test and within the LS cohort using the pairwise Mann–Whitney U test.

For beta diversity estimation, any dissimilarity measure can be used, with Bray-Curtis dissimilarity commonly used in microbiome studies to consider the compositional nature of relative abundance

data. Bray–Curtis dissimilarity were calculated from genus level taxonomic profiles ("vegdist" function, method = "bray," R vegan). Principal coordinate analysis (PCoA) was performed on the calculated Bray–Curtis dissimilarity to visualize gut microbiota composition ("pcoa" function, R ape).

Quantification of the variance within the gut microbiota composition was evaluated with permutational multivariate analysis of variance (PERMANOVA) ("adonis2" function, permutations = 10000, by = "margin," R vegan). Demographics (e.g., age and sex) and medical information (e.g., endoscopy findings, mutated MMR genes, and colectomy records) were selected as covariates to explain the taxonomic profile variances at genus level.

Subjects were clustered based on gut microbiota composition to identify potential LS-specific enterotypes[125,126] by fitting all genus-level abundances into a Dirichlet Multinomial Mixture (DMM) model ("dmn" function, R DirichletMultinomial)[127]. The optimal number of community types was determined based on the lowest Laplace approximation score.

2.8.2 Differential abundance analysis

Microbial features (species and KOs) with low read alignment or coverage and low prevalence were discarded (read alignment cutoffs for species and read coverage cutoffs for KO were 10.0, respectively; the prevalence cutoff was 5%). Similarly, fecal metabolites with low concentrations and prevalence were discarded (concentration cutoff: 10 nmol/g; prevalence cutoff: 20%). A higher prevalence threshold for fecal metabolites was used to remove batch-specific metabolites.

Pairwise Mann–Whitney *U* test ("wilcox_test" function, R rstatix) was performed within the LS cohort or sporadic cohort to identify taxonomic, metagenomic and metabolite features with differential abundances. The Mann–Whitney *U* test was also performed between the LS and non-LS controls. *P* value of 0.05 was considered at be statistically significant.

The generalized fold-change, a pseudo-fold-change calculated as the geometric mean of the differences between quantiles across both groups, was calculated for each comparison (adapted from R-SIAMCAT[128]) to quantify difference of abundance. Effect size estimation and magnitude interpretation was done for each comparison ("wilcox_effsize" function, R rstatix). Benjamini–Hochberg correction, a false discovery rate estimation method, was performed on the *P* values to obtain the FDR corrected *q* value ("p.adjust" function, method = "BH," R stats).

2.8.3 Trend test

Previous studies have reported several dysbiosis patterns across sporadic CRC's adenoma-carcinoma model[42]. For example, *Atopobium parvulum* was enriched in colorectal adenoma and intramucosal carcinoma subjects, but no longer enriched in CRC patients. In contrast, *Fusobacterium nucleatum* was consistently enriched along adenoma-carcinoma model, with

depleted abundance in post-surgery patients. Other CRC-associated microbiota, such as *Parvimonas micra* and *Peptostreptococcus stomatis* were enriched only in late-stage CRC patients. Different dysbiosis patterns might be associated to the causal roles of CRC-associated microbiota, with those enriched in early stage of CRC progression being more likely to affect CRC progression. Here, Jonckheere-Terpstra trend test was utilized to detect microbiota with specific enrichment trends across adenoma-carcinoma model.

As Jonckheere-Terpstra test only detects increasing or decreasing trend across specific groups, two-group comparison (Mann-Whitney U test) across the groups was performed to specify when the potential dysbiosis occurs. For the LS cohort, dysbiosis trend between LS control to LS CRC and an additional two-group comparison between LS CRC vs LS surgery was explored. For the sporadic CRC cohort, dysbiosis trend between healthy controls to late-stage CRC (CRC Stage III and Stage IV) was explored.

2.8.4 Association test

Several other host factors such as age, sex, and BMI have been associated with the human gut microbiota[5,129]. As the cohort utilized in this study was recruited in a natural manner, such host factors might be detected as LS-microbiota association. Thus, it is essential to detect other associations to extract robust LS-microbiota association. Here, a generalized linear model implemented in MaAsLin2[130] R package were utilized to consider the effect of other host factors in LS-microbiota association.

2.8.5 Classifier training, hyper-parameter tuning and evaluation

Statistical test detection power was limited by low sample numbers. To validate the findings identified by the statistical tests, machine learning-based marker screening was performed as a stricter standard[6–8,15,131]. As previous studies have shown that gut microbiota profiles vary according to CRC progression stage, five classification tasks were used: differentiation between non-LS controls and 1) sporadic adenoma, 2) stage 0 CRC, 3) stages I/II CRC, 4) stages III/IV CRC, and 5) stages I–IV CRC. Only non-LS subject-species profiles were used to train the classifiers as the LS cohort was too small for classifier training. Feature contribution to model performance was estimated, with top contributing features picked as putative CRC-associated microbes.

Abundance filtering was performed prior to training by selecting species with a mean abundance higher than a specific threshold for classifier training. Three commonly used classifier architectures (logistic regression, Random Forest and gradient boosted decision tree) were used, with the best model for each task chosen as the final model. Bayesian optimization was performed in a nested cross-validation (CV) framework for hyperparameter tuning. Five inner loops ("BayesSearchCV" function, n_iter = 30, Python scikit-optimize) and five outer loops ("cross_validate" function, scoring= "roc_auc," Python scikit-learn) were used for the nested CV.

Data split was done in a stratified fashion for both inner and outer fold to ensure similar case and control balance ("StratifiedKFold" function, n_splits = 5, shuffle = True, random_state = 0, Python scikit-learn). Details of the filtering thresholds, classifier architecture, and hyperparameter search ranges are listed in **Appendix 1**.

Classifier performance was evaluated by area under the receiver-operating characteristic curve (AU-ROC) ("roc_curve" function, Python scikit-learn). Classifier performance during training was evaluated using the mean AU-ROC of the five-fold CV. The classifier performance for the three LS subject classification tasks (LS control vs. LS-adenoma, LS control vs. LS-CRC, and LS control vs. LS-colectomy) was evaluated with no CV.

CRC probability of each subject was estimated using scikit-learn "predict_proba" function. For non-LS subjects, CRC probability estimation was performed in a leave-one-out manner ("LeaveOneOut" function, Python scikit-learn), with classifier parameters set to be identical with the final classifier. For the patients with LS, CRC probability estimation was parallelly performed with classifier performance evaluation. Because the CRC probability may be affected by the classifier performance and class ratio, the estimated probability values was normalized with a min–max normalization. Probability normalization was performed within each classification task.

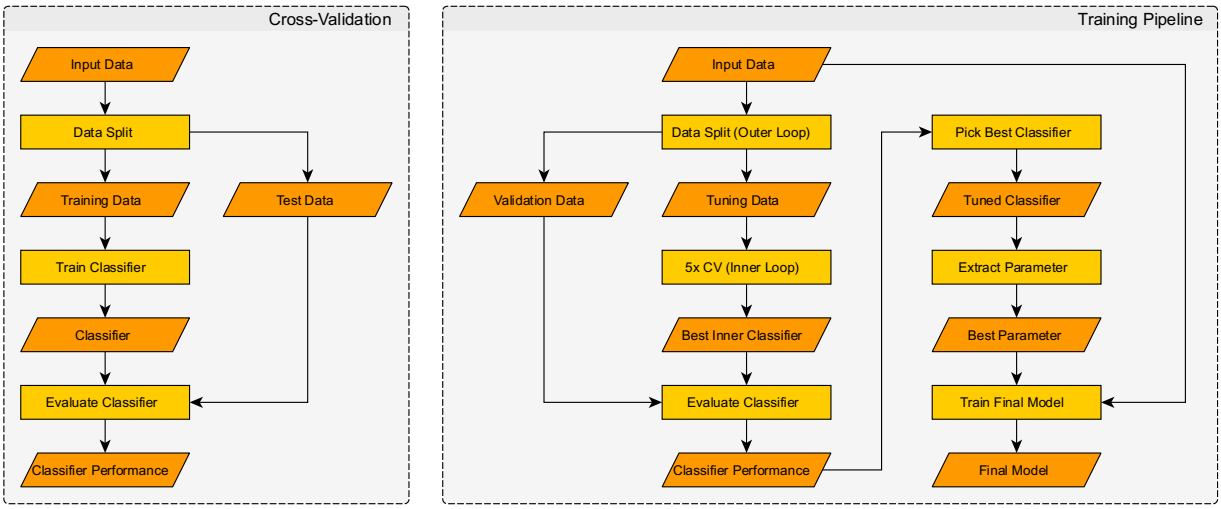


Figure 2.4 CRC classifier training pipeline

Classifier training was performed in a nested CV framework for hyper-parameter tuning, with 5x CV done both in inner and outer loop. Final classifier was then utilized in downstream importance analysis and LS CRC classification task. Cross-validation was performed by splitting input data into training and test data, in order to evaluate classifier performance during training.

2.8.6 Local feature importance analysis

Previous studies have shown gut microbiota profile can be utilized to differentiate CRC patients from non-tumor control[15,131,132]. Furthermore, feature selection and contribution analysis have

been utilized to screen for screen for potential CRC microbial markers. In this study, species importance was obtained using "feature_importance" function (parameter importance = "gain") from Python LightGBM. Species contributions obtained in this manner were estimated for the entire model (global importance) rather than for specific subject prediction (local importance).

A previous study proposed that while global importance is sufficient to obtain a general overview of microbe-CRC associations, local importance estimation is essential to elucidate microbe–CRC associations for a specific subset of subjects[133]. Here, "TreeExplainer" function of Python SHAP[134] was used to estimate Shapley Additive Explanations (SHAP) values. The estimated SHAP values served as indicators of the local importance of each species in CRC prediction in individual subjects.

Principal coordinate analysis (PCoA) was performed on the estimated SHAP values for data exploration ("prcomp" function, R stats). To identify potential LS CRC-specific clusters, k-means clustering was performed on the first two principal components (PC1 and PC2) of correctly predicted LS and non-LS CRC subject's SHAP value ("kmeans" function, R 'stats'). The optimal number of clusters were decided by an elbow plot of the total within cluster square sums ("fviz_nbclust" function, FUNcluster = 'kmeans,' method = 'wss,' R factoextra).

To obtain a visual representation of the local feature importance for LS CRC classification and the feature importance trends within a specific cluster, the correctly predicted SHAP values of the LS CRC subjects SHAP values was summarized into a waterfall plot.

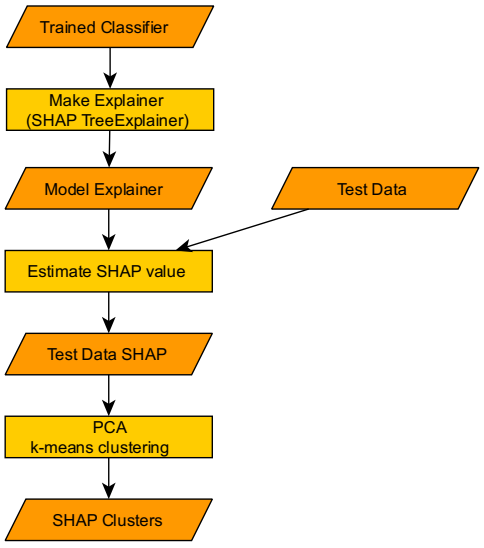


Figure 2.5 SHAP analysis pipeline

SHAP explainer was built for the final model obtained from **Figure 2.4** training, which was then used to estimate SHAP values of test data (LS subject's species-level profile). In order to visualize and identify distinct subject's cluster, PCA was done on estimated SHAP values, followed with k-means clustering on the principal components.

2.8.7 Over-representation analysis

Overrepresentation analysis was performed to identify metabolic modules with an overrepresentation of KOs and metabolites that were differentially abundant between the case and control groups. The feature set used for ORA was created by aggregating both KOs and KEGG Compounds into KEGG MODULE level metabolic modules (KEGG database obtained on 2023/03/31). Furthermore, to encompass metabolic modules not available in KEGG, another custom gene set was created by aggregating both KOs and KEGG Compounds into EnteroPathway[135] MODULE level metabolic modules (EnteroPathway database obtained on 2023/10/05) (**Appendix 2**). ORA was performed as implemented by the "enrichr" function of R clusterProfiler. Multiple testing corrections were performed using the Benjamini–Hochberg method, and modules with FDR $P < 0.1$ were considered as overrepresented modules.

Chapter 3
Results and Discussion
3.1 Subject overview

Based on statistical test on subject's demographic (**Figure 3.1**), LS controls were significantly younger than LS adenoma and LS surgery. Although not significant, LS controls were younger compared to LS CRC, as expected based on the early onset of tumors in LS subjects. LS-CRC had a higher male-to-female ratio, smoker-to-non-smoker ratio, and Brinkman index than LS control, with only Brinkman Index showing statistical significance ($P > 0.05$) (**Figure 3.1**). No significant differences were observed between patients with LS in mutated MMR genes and dietary fiber or meat intake. Similar to the LS cohort, the controls in the sporadic cohort were significantly younger than those with CRC progression. In addition, the Brinkman Index was significantly higher in patients with CRC progression (**Figure 3.1**). Overall, clinical data showed putative association between smoking habit and CRC in both LS and sporadic cohort.

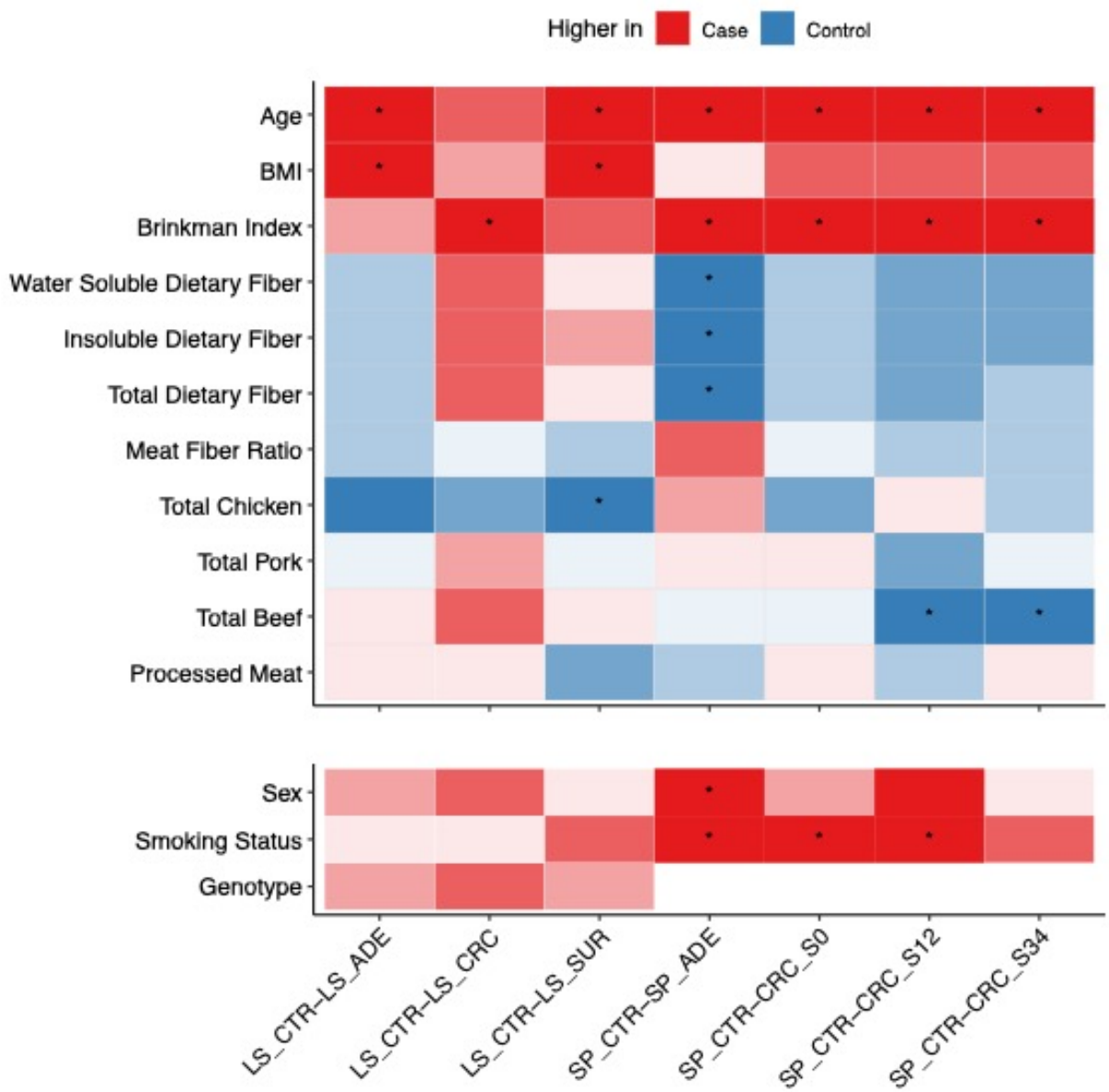


Figure 3.1 Covariate analysis results

Heatmap of subject's demographic comparison results. Upper heatmap was those of numerical variables, which were compared with non-parametric Wilcoxon-Mann-Whitney U test. Lower heatmap was categorical variables, compared with Chi-square test. Color indicate in which group was the covariate higher in, control vs case was shown in the x-axis.

3.2 Lynch Syndrome subjects have distinct gut microbiota composition

3.2.1 LS subjects gut microbiota has lower alpha diversity than non-LS subjects

Species alpha diversity, estimated using the Shannon–Wiener index, was significantly lower in the LS cohort compared to the sporadic CRC cohort (Mann–Whitney U $P = 0.00014$, **Figure 3.2A**).

Lower alpha diversity in LS subjects compared to sporadic CRC and non-LS controls was in line with previous studies, suggesting distinct LS gut microbiota[92–94].

Pairwise comparisons within the LS cohort did not detect any significant association between alpha diversity and patient group, MMR mutations, or age (**Figure 3.2B**). Although not significant, MSH2+EPCAM had low alpha diversity even within the LS cohort (**Figure 3.2B, middle)**. This observation indicates LS distinct gut microbiota might be formed prior to adenoma or carcinoma formation, although a larger cohort is essential to validate this finding.

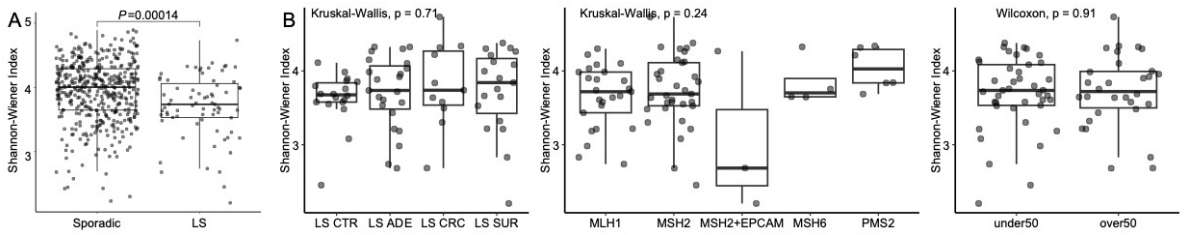


Figure 3.2 Gut microbiota alpha diversity comparison results

Boxplots of Shannon-Wiener index. A) Sporadic vs LS comparison showed LS alpha-diversity significantly lower than sporadic cohort. B) No significant difference of alpha diversity was detected when LS subjects were compared between colonoscopy findings (left), MMR mutation (middle), and age (right).

3.2.2 CRC and colectomy status has stronger association with gut microbiota composition than LS status

Principal coordinate analysis (PCoA) was performed on Bray–Curtis dissimilarities of the patient's genus-level profile to visualize the gut microbiota composition. In the PCoA plot, no clear separation was observed between the patients with LS and those with sporadic CRC (**Figure 3.3A**). Focusing on the gut microbiota profile of the groups, the PCoA plot showed no clear separation between the LS control, LS-adenoma, LS-CRC, and LS-surgery groups (**Figure 3.3B**).

Permutational analysis of variance (PERMANOVA) was performed to verify whether LS affected the gut microbiota composition. Colonoscopy finding (four LS groups and five sporadic CRC groups) was found to be the best explanatory variable (R^2 = 1.885%, P = 0.161). Of interest, LS diagnosis was the second-best explanatory variable and showed statistical significance (R^2 = 0.328%, P = 0.110) (**Figure 3.3C, left**). This is in line with alpha-diversity results, indicating potential distinct LS gut microbiota composition. However, statistical significance was not observed for PERMANOVA with only non-LS controls and LS controls, suggesting subject's CRC progression as confounding factor[94,136]. Thus, a larger LS cohort with non-tumor controls is essential for a more robust analysis.

PERMANOVA within the LS cohort showed that the best explanatory variables were MMR gene mutations ($R^2 = 5.403\%$, $P = 0.514$), followed by sex, age, colonoscopy diagnosis, and colectomy record (**Figure 3.3C, middle**).

Based on PERMANOVA results, previous reports of LS cohort distinct gut microbiota might be confounded by subject's clinical history, such as past adenoma formation or surgery. A previous study has reported LS status were not explanatory of gut microbiota composition when adenoma or carcinoma was accounted for[94]. PERMANOVA within the LS cohort showed that MMR mutations, which is associated with CRC risk in LS subjects[70], were the most explanatory variables for gut microbiota variance. In order to deconfound MMR mutation effect on CRC-gut microbiota association study, a larger cohort or a specific mutation cohort is essential. Finally, although colectomy history was less explanatory compared to other variables (**Figure 3.3C, middle**), colectomy parts were the most explanatory variables for those with colectomy records, in line with a previous study[95].

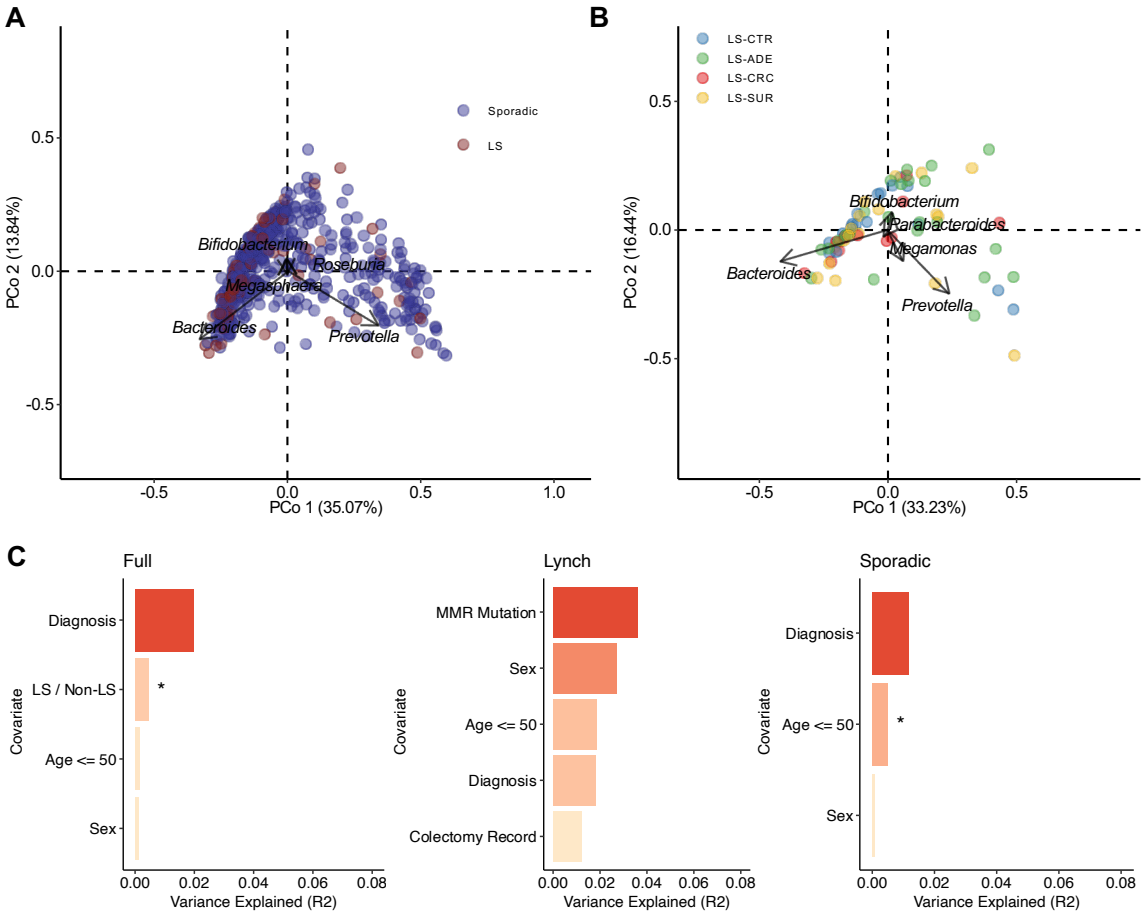


Figure 3.3 PCoA and PERMANOVA results

A) Scatterplot of all subject's principal coordinates, annotated by LS status. B) Scatterplot of LS subject's principal coordinates, annotated by colonoscopy findings. C) Barplot of PERMANOVA results done on the full cohort (LS + non-LS, left), LS (middle), and non-LS (right); x-axis represents gut microbiota composition variance explained by covariates. Asterisk represents statistical significance ($P < 0.05$).

3.2.3 LS subjects tend to have Faecalibacterium-poor gut microbiota

To check whether LS-associated enterotypes existed, the gut microbiota profile of each patient was fitted into a Dirichlet Multinomial Mixture (DMM) model[127]. From the best-fit model, four enterotypes were determined, consisting of three *Bacteroides*-dominant enterotypes (Enterotypes 1, 2, and 3) and one *Prevotella*-dominant enterotype (Enterotype 4) (**Figure 3.4A-C**). Notably, Enterotypes 1 and 4 were characterized by a higher proportion of *Faecalibacterium* than the other two enterotypes (**Figure 3.4C**). Comparison of enterotype ratios between the LS and sporadic cohorts showed that the LS cohort had a significantly higher ratio of Enterotype 2 (Fisher's exact test $P = 8.50 \times 10^{-4}$) and a significantly lower ratio of Enterotypes 1 ($P = 0.0102$) and 4 ($P = 0.0139$) compared to the sporadic cohort (**Figure 3.4D**). However, *Faecalibacterium* abundance in the LS and sporadic cohorts was not significantly different.

Based on enterotype analysis, patients with LS tended to cluster into enterotypes with high *Bacteroides* abundance and low *Faecalibacterium* abundance (Enterotypes 2 and 3). In comparison, non-LS patients tended to have a higher prevalence of enterotypes with a high *Faecalibacterium* abundance (Enterotypes 1 and 4). Enterotypes 2 and 3 proportions were similar across the LS control, LS-adenoma, LS-CRC, and LS-surgery, indicating a minimal association between adenoma and carcinoma formation or colectomy to enterotype. Low alpha diversity and enterotypes with high *Bacteroides* and low *Faecalibacterium* counts have been associated with IBD[11,137]. Thus, low alpha diversity observed in section 3.2.1 and low *Faecalibacterium* enterotypes of LS subjects with LS indicated intestinal inflammation. Furthermore, activated immune profiles have been reported in patients with LS across all stages of CRC progressions, including prior to adenoma formation (LS control)[79,81,138,139].

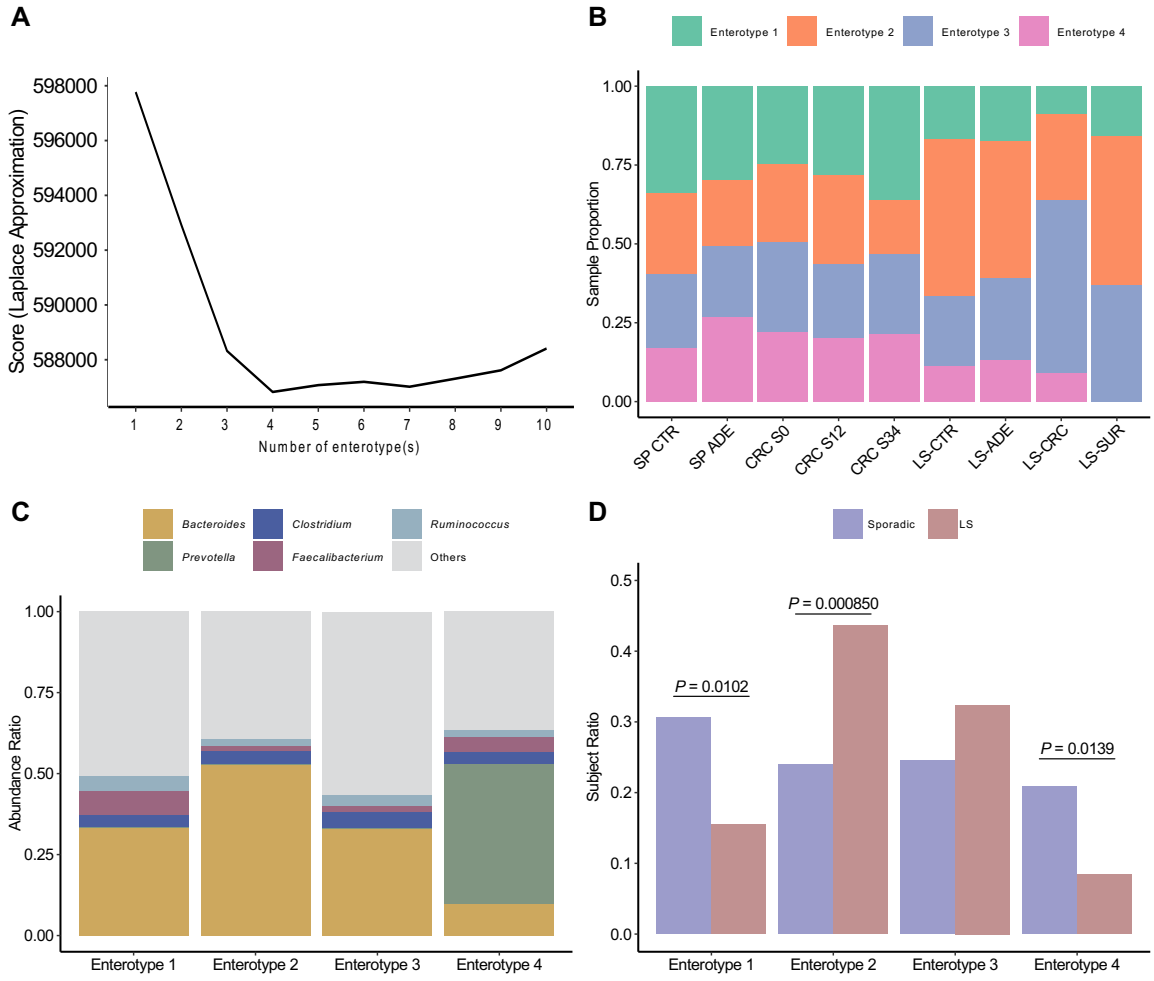


Figure 3.4 Enterotype analysis results

A) Elbow plot of goodness-of-fit (estimated by Laplace approximation) of the input data to DMM model for varying mixture (enterotypes) number. B) Subject's proportion of defined enterotype within each subject's group. C) Dominant genus of each enterotypes. D) Comparison between enterotype proportion in non-LS and LS. Comparison was done with Chi-square test.

3.3 Distinct *Fusobacterium* species enrichment in the gut microbiota of LS CRC patients

3.3.1 CRC-associated microbiota, except for Fusobacterium, had non-significant change in LS CRC

To identify the gut microbial species that were differentially abundant between dysbiosis occurring in patients with and without LS after adenoma and CRC formation in more detail, I performed a differential abundance analysis within non-LS and LS cohorts.

A pairwise comparison of species abundances between the LS controls and other groups detected 103 species with differential abundances (Mann–Whitney U, $P < 0.05$) in at least one comparison

(**Figure 3.5**, **Appendix 3**). Focusing on the differentially abundant species between the LS control and LS-CRC groups, 19 were enriched, and 29 were depleted in the LS-CRC group. Of the 19 species enriched in LS-CRC, 12 belonged to the *Fusobacterium* genus, such as *F. nucleatum* subsp. *nucleatum* and *F. nucleatum* subsp. *animalis*. Other CRC-related species, such as *Peptostreptococcus stomatis*, *P. anaerobius*, *Parvimonas micra*, and *Gemella morbillorum*, were not significantly enriched in LS-CRC.

Results indicated a distinct association between *Fusobacterium* species and LS-CRC, which might be explained by the early stages (Stages I and II) of LS CRC, as other sporadic CRC-related species were more associated with advanced stages (Stages III and IV)[6,42]. In line with a previous study, *F. nucleatum* enrichment was not observed in patients with LS and adenoma was not found[95].

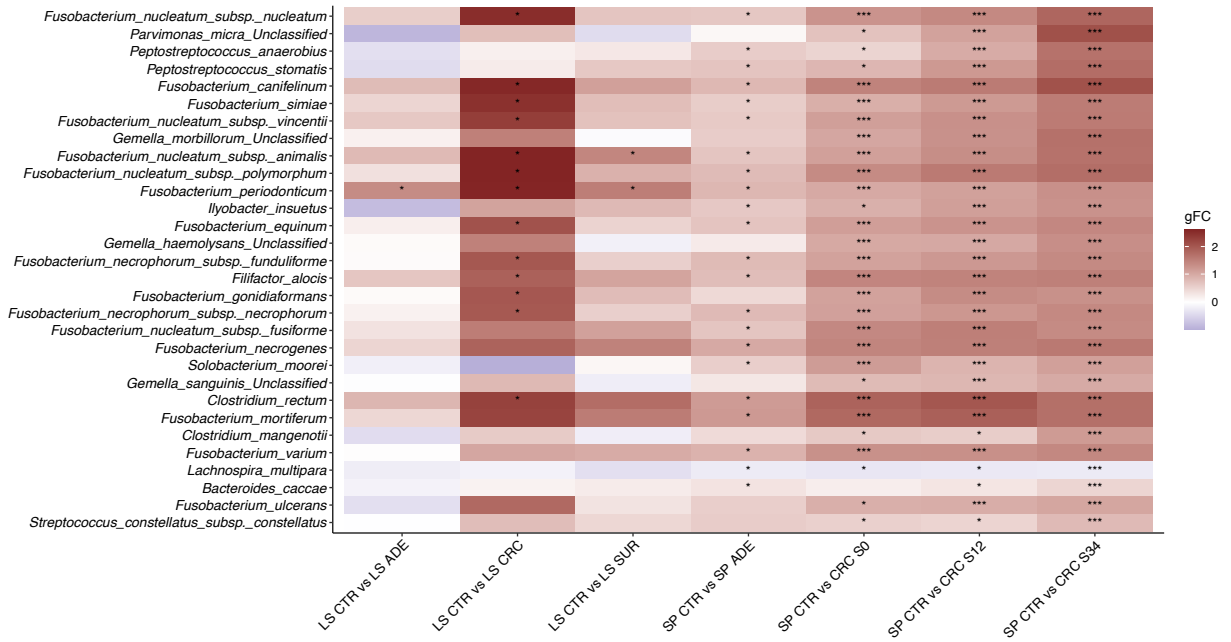


Figure 3.5 Top 20 statistically significant CRC-associated microbiota

Twenty features with the lowest p-value (Wilcoxon-Mann-Whitney U test) in non-LS control vs CRC S34 was picked for visualization. Features were arranged by p-value in ascending order (lowest to highest). Asterisks indicate statistical significance (* $P < 0.05$, *** FDR $P < 0.1$). Color indicates feature abundance differences, estimated by generalized fold-change; red indicate features enriched in case group and blue indicate features enriched in control group. Control vs case was shown in x-axis.

3.3.2 Fusobacterium enrichment occurs only in LS CRC patients, not in LS adenoma and LS surgery

Previous study has shown the association of gut microbiota and CRC progression, with dysbiosis occurring earlier or later in the adenoma-carcinoma model[42]. For example, *Atopobium parvulum* was enriched in colorectal adenoma patients, but not in CRC patients. In contrast, *F. nucleatum* was consistently enriched along CRC progression and *P. stomatis* was especially enriched in late-

stage CRC. Furthermore, a previous study reported no *Fusobacterium* enrichment in LS subjects with colorectal adenoma. Thus, trend test on species profile was performed to identify potential LS-specific dysbiosis pattern.

Jonckheere-Terpstra test results showed *Fusobacterium* species abundance has significant increasing trend across adenoma-carcinoma model (**Figure 3.6**). Pairwise Mann-Whitney test further showed that significant enrichment was only seen in LS adenoma vs LS CRC comparison, but not in LS control vs LS adenoma comparison. Next, LS CRC vs LS surgery comparison showed significant depletion of *Fusobacterium* abundance in LS surgery. Furthermore, Jonckheere-Terpstra test did not detect significant increasing trend in other CRC-associated microbiota, suggesting specific enrichment of *Fusobacterium* species in LS CRC. Thus, trend test results further support previous observation of enriched *Fusobacterium* species and that the enrichment was specifically significant only in LS CRC subjects.

Notably, *Fusobacterium* enrichment pattern in LS CRC was distinct compared to sporadic CRC dysbiosis pattern (**Figure 3.7**). Specifically, the most *Fusobacterium* spp. prominent enrichment in LS CRC was between LS adenoma and LS CRC. In comparison, the most prominent *Fusobacterium* spp. enrichment in sporadic CRC was between non-LS control and sporadic adenoma patients.

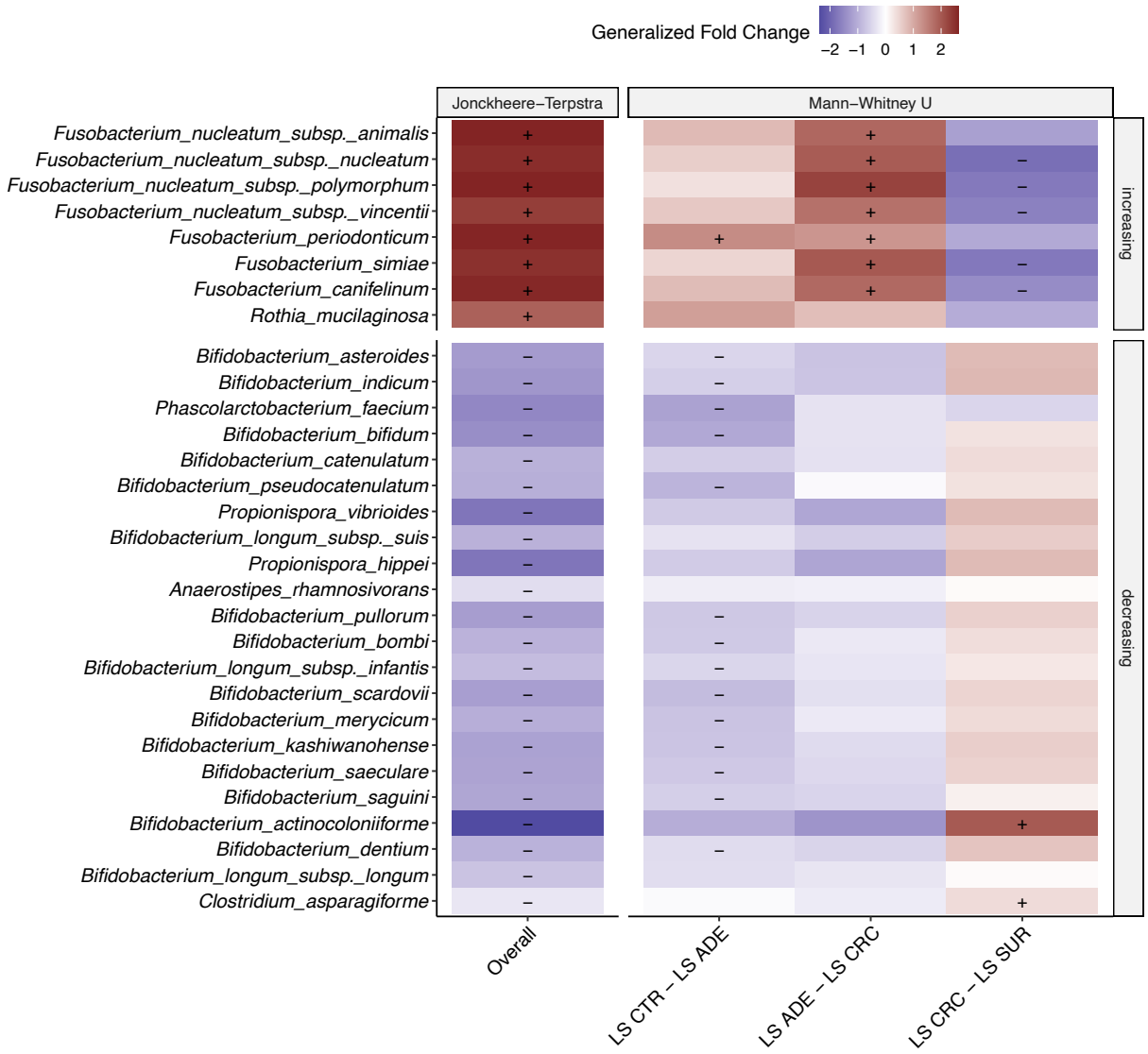


Figure 3.6 Top 30 features with significant dysbiosis trend in LS cohort

Heatmap indicate overall dysbiosis pattern trends (from LS control to LS CRC) and dysbiosis pattern between CRC progression stage, splitted into increasing (upper-half) and decreasing pattern (bottom-half). Features arranged based on the p-values of overall trend in ascending order. Positive/negative sign indicate statistical significance ($P < 0.05$). Color indicates feature abundance differences, estimated by generalized fold-change; red indicate features enriched in case group and blue indicate features enriched in control group. Control vs case was shown in x-axis.

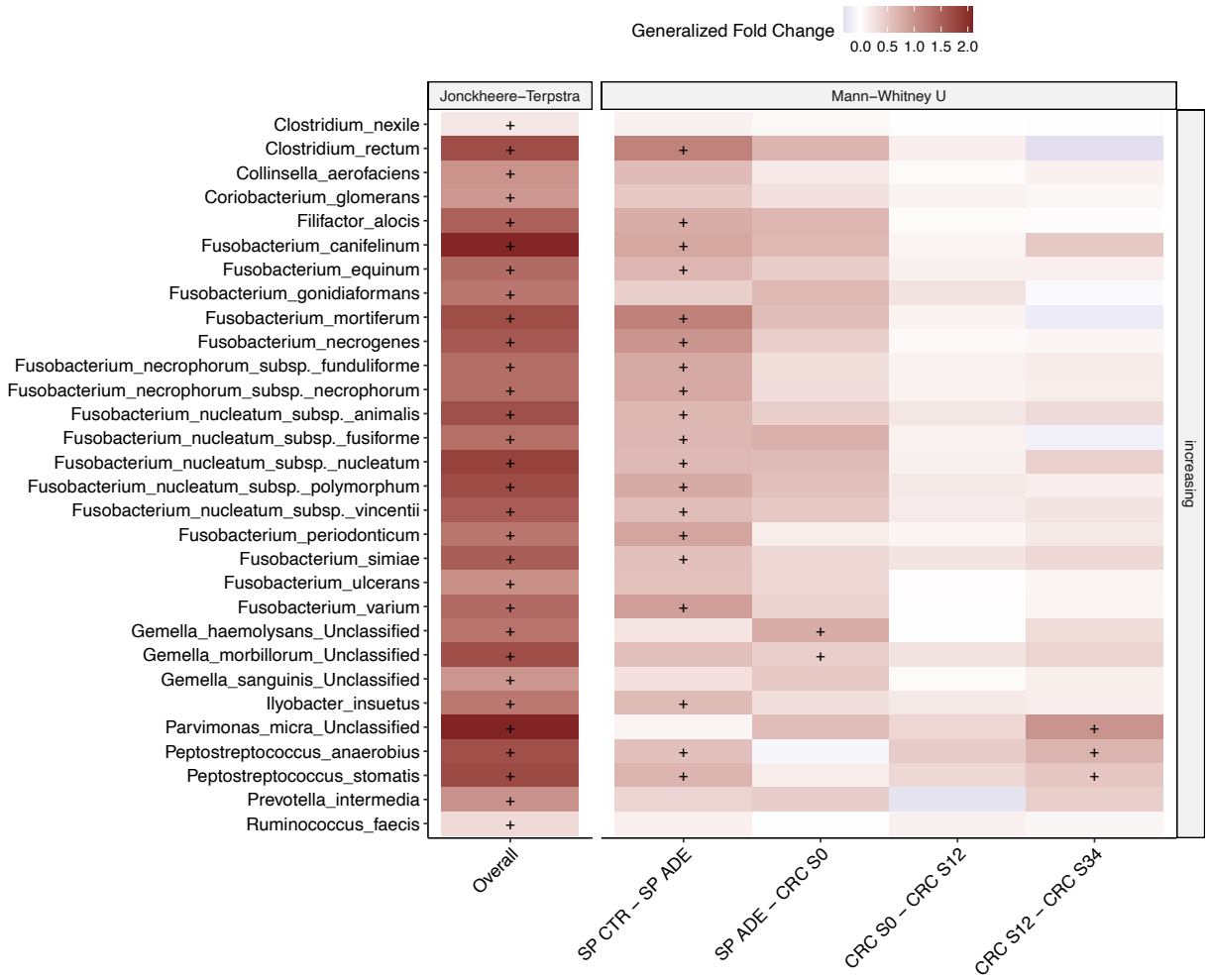


Figure 3.7 Top 30 features with significant dysbiosis trend in sporadic CRC cohort

Heatmap indicate overall dysbiosis pattern trends (from non-LS control to CRC S34) and dysbiosis pattern between CRC progression stage. Features arranged based on the p-values of overall trend in ascending order (top 30 features all showed increasing trend). Positive sign indicates statistical significance ($P < 0.05$). Color indicates feature abundance differences, estimated by generalized fold-change; red indicate features enriched in case group and blue indicate features enriched in control group. Control vs case was shown in x-axis.

3.3.3 Confounding factors has minimal effect on CRC patients Fusobacterium enrichment

Previous studies have reported on several potential factors confounding microbiota-disease association, such as age and sex[5,129]. As the LS controls subjects were significantly younger than the rest of the LS cohort, I explored whether other variables was associated with *Fusobacterium* enrichment by utilizing a multi-variate linear model-based association test (MaAsLin2)[130].

MaAsLin2 results for LTP profile showed *Fusobacterium* enrichment was associated only with carcinoma progression ($q < 0.25$), as no significant association with age, sex, colectomy status, BMI and smoking status was detected (**Figure 3.8**). Of interest, subjects with PMS2 variants

showed negative association with *Fusobacterium* species abundance, with PMS2 variants reported to have lower CRC risk compared to LS subjects with MLH1 or MSH2 mutations[140,141].

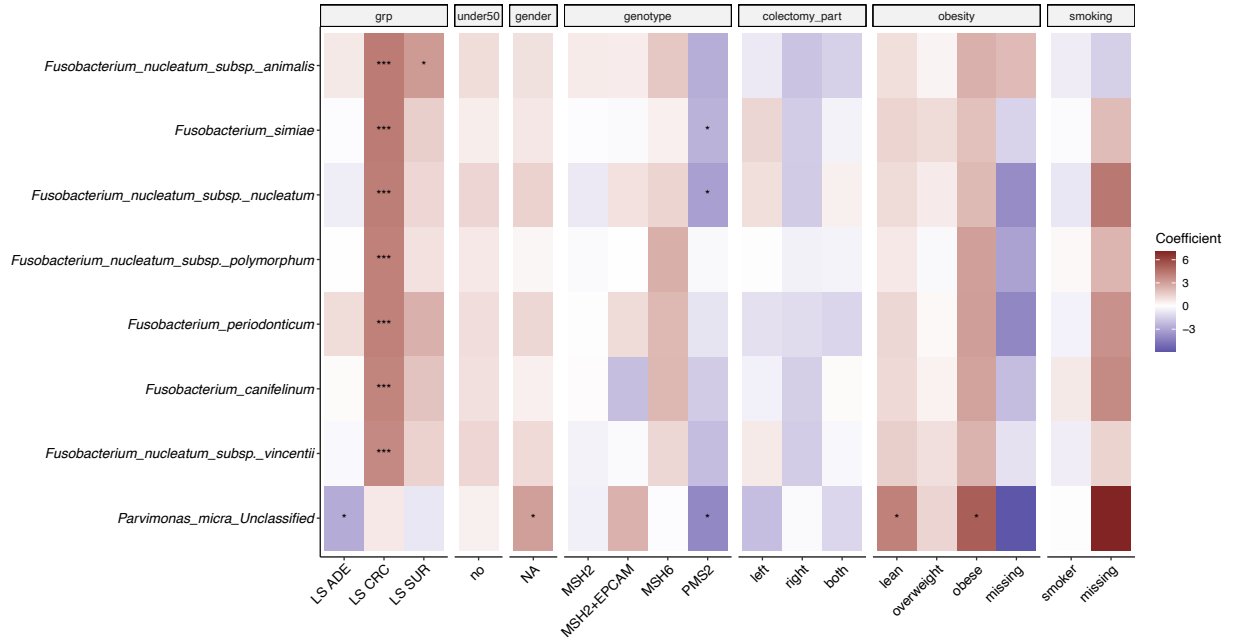


Figure 3.8 MaAsLin2 results of CRC-associated microbiota enriched in LS CRC

Association test results filtered for CRC-associated microbiota, as shown in **Figure 3.5**. Asterisks indicate statistical significance (* $P < 0.05$, *** FDR $P < 0.25$; the default threshold for MaAsLin2). Color indicates feature abundance differences, estimated by linear model coefficient; red indicate features enriched in case group and blue indicate features enriched in control group. Control group for each tested covariate was LS control (colonoscopy finding), MLH1 (MMR mutation), none (colectomy part), normal (obesity), non-smoker (smoking status).

3.4 Machine learning-based feature selection validated *Fusobacterium* species enrichment as an LS-CRC marker

3.4.1 Gut microbiota profile-based CRC classifier capable of differentiating CRC patients and non-tumor controls

Differential abundance analysis on taxon profiles suggests *Fusobacterium* spp. enrichment is a shared CRC-related dysbiosis across LS and sporadic CRC. However, the test has limited detection power due to the LS cohort's small sample size. In addition, the results are not necessarily reproducible in other taxonomic annotation pipelines. Finally, no previous studies using whole-genome shotgun sequences of LS-CRC is available for external validation at the time of writing.

Nevertheless, accumulated knowledge has been obtained from a considerable amount of microbial data on sporadic CRC[6–8]. One commonly used method is to build a machine learning-based classifier to screen for microbial markers. Here, classifiers were trained to differentiate CRC patients and non-tumor control based on their gut microbiota taxon profiles to screen for CRC-

associated gut microbiota. LS cohort data was used as test data to explore potential resemblance between microbial markers across sporadic and LS CRC.

Models were trained as described in section 2.8.5. Based on model performance, gradient-boosted decision tree (GBDT) classifier showed the best performance to differentiate between non-LS controls and patients with sporadic CRC based on their species profiles. The trained classifier performed comparable to previous CRC studies[6–8], as evaluated by its mean area under the receiver operating curve (AU-ROC) across five-fold cross-validation (mean AU-ROC = 0.805; **Figure 3.9A**). Feature importance analysis showed that the top contributors were mostly CRC-related species, such as *F. nucleatum* subsp. *nucleatum*, *Gemella morbillorum*, *P. stomatis*, and *P. micra* (**Figure 3.9D**).

3.4.2 CRC classifier trained on non-LS subjects can differentiate LS CRC patients and LS controls

The constructed sporadic CRC classifier was tested for its ability to discriminate LS CRC from LS controls using the species profiles of patients with LS as the input. The model distinguished LS-adenoma, LS-CRC, and LS-surgery from LS control with AU-ROC of 0.575, 0.869, and 0.684, respectively (**Figure 3.9B**). The normalized CRC probability predicted by the classifier was higher for LS-CRC and lower for LS controls, whereas LS-adenoma and LS-surgery showed a more uniform distribution (**Figure 3.9C**). These results indicate some overlap between sporadic CRC and LS CRC microbial markers, which were specific to CRC subjects.

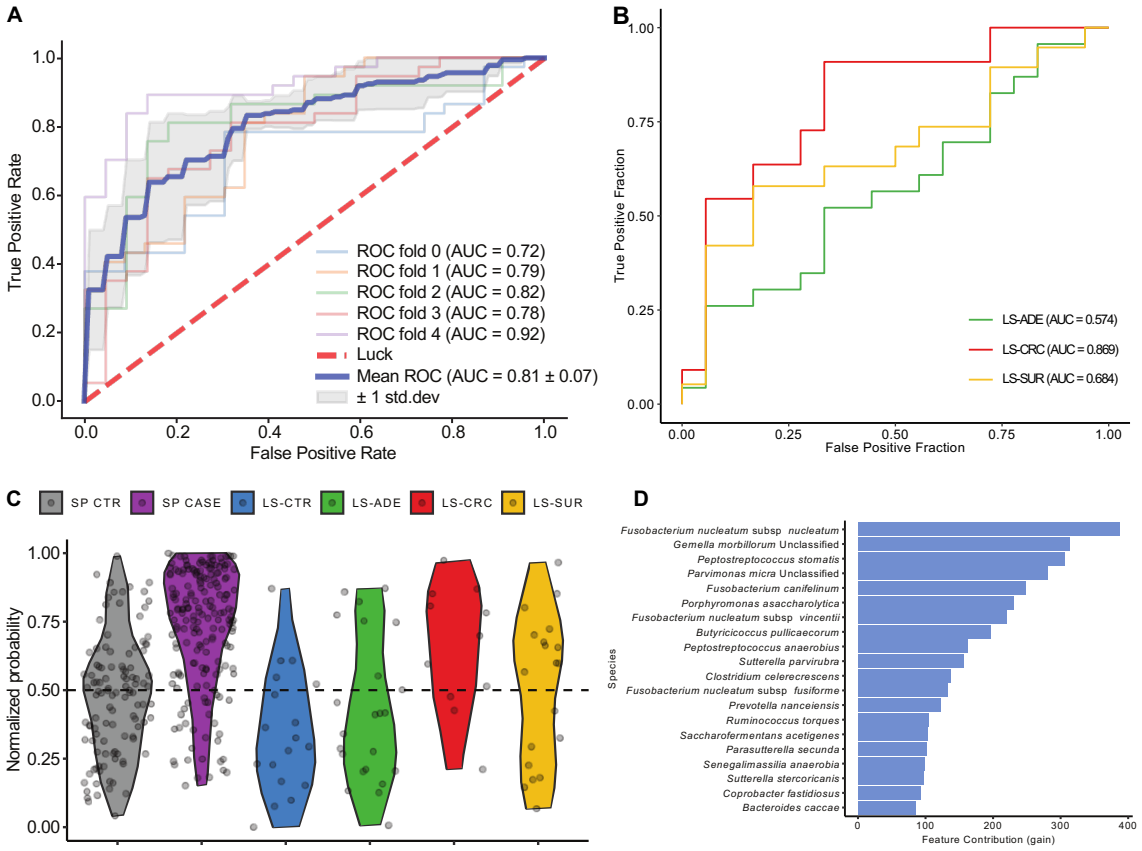


Figure 3.9 Gut microbiota-based CRC classifier

A) Classifier performance during training estimated by AU-ROC, thin lines represent performance within each CV fold, thick line represents mean AU-ROC. B) Classifier performance for each LS classification tasks; which is to differentiate LS control and a case group (one of LS adenoma, LS CRC, or LS surgery). C) subject's CRC probabilities predicted by classifier, normalized for each task. D) Feature contribution to classifier performance, estimated by performance gained by utilizing said features.

3.4.3 F. nucleatum subsp. nucleatum identified as top microbial marker for LS CRC

The classifier performance and feature importance analysis results indicated that CRC-related species abundance was the main predictor for differentiating between LS-CRC and LS controls. However, it was not sufficient to screen for LS CRC markers because feature importance was estimated for the entire classifier (global importance), which was trained on sporadic species profiles[133].

A previous study proposed a framework utilizing the Shapley Additive Explanations (SHAP) value to estimate feature importance for specific subjects (local importance) and showed distinct clusters within patients with sporadic CRC based on the estimated local importance[133]. Here, I used the

SHAP value-based local feature importance analysis framework to validate whether *Fusobacterium* spp. were the main predictors of LS-CRC.

The PCA of the estimated SHAP values showed a distinct separation between control subjects and patients with CRC (**Figure 3.10A**). Next, the SHAP value principal components of correctly predicted patients with CRC were clustered via k-means clustering (optimal k = 4, determined by an elbow plot). Of the four clusters, only Clusters 2 and 4 contained subjects with LS-CRC (**Figure 3.10B**). In contrast, patients with sporadic CRC spread over the four clusters.

From each cluster's LS-CRC patient SHAP value, I found that the top-ranking species in Cluster 2 were CRC-related microbiota, such as *G. morbillorum*, *F. nucleatum* subsp. *nucleatum*, *P. micra*, and *P. stomatis* (**Figure 3.10C**). The top-ranking species in Cluster 4 were similar to those in Cluster 2, with a highly abundant *F. nucleatum* subsp. *nucleatum* as the top contributor (**Figure 3.10D**). Overall, the local feature importance analysis suggests that the sporadic CRC classifier distinguished LS-CRC from LS control, mainly based on *Fusobacterium* spp. presence, especially *F. nucleatum* subsp. *nucleatum*. Therefore, ML-based microbial marker screening results supports previous observations in differential abundance analysis and trend test of *Fusobacterium* species associations with LS CRC.

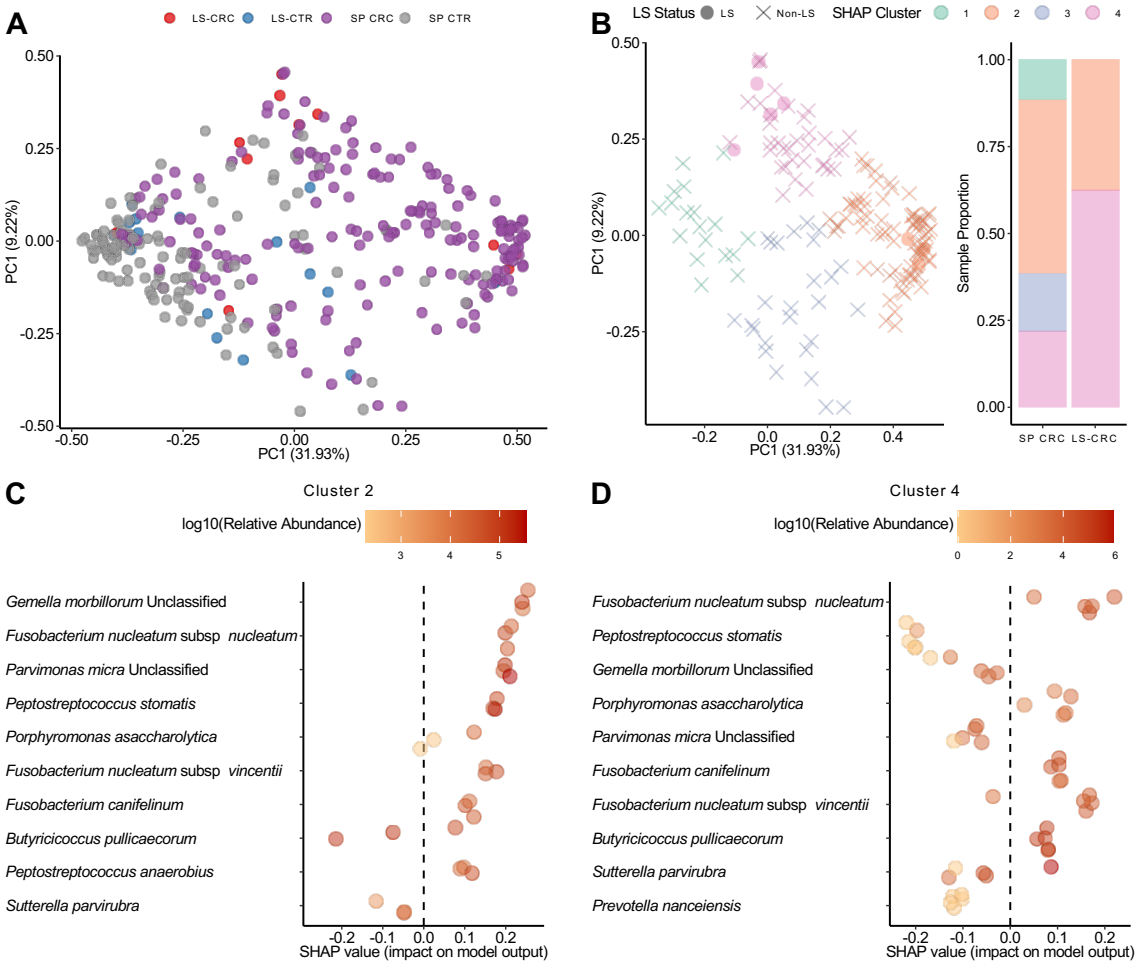


Figure 3.10 SHAP-based local feature importance analysis results

A) Scatterplot of LS and non-LS controls and CRC subject's SHAP value principal components, annotated by LS and CRC status. B) Scatterplot of LS and non-LS CRC subjects correctly predicted by classifier, annotated by clusters defined with k-means clustering; subject's proportion for each cluster shown in stacked bar plot in right panel. C) Waterfall plot of cluster 2 LS CRC subject's local importance, x-axis represents feature's SHAP value (positive means CRC-associated), color represents feature abundance in subject. D) Waterfall plot of cluster 4 LS CRC subject's local importance.

3.5 Distinct enrichment pattern of bacterial virulence factors in LS-CRC subjects

3.5.1 Enriched fap2 in LS-CRC subjects

F. nucleatum genes have been reported as virulence factors (Fn VF) associated with CRC progression, such as *fadA*, *fap2*, *cbpF*, *radD,* and *FnDps*[31,48,142–146]. Here, I explored the LS metagenome data to identify the potential enrichment of these virulence factors.

Four out of seven Fn VFs were detected in our LS cohort, which were *fap2* (average prevalence 5.0%), *radD* (average prevalence 35.7%), *fadA* (average prevalence 7.1%), and *cbpF* (average prevalence 10.4%), while *fadA2*, *fadA3*, and *FnDps* were not detected in our LS cohort (**Figure 3.11A**). Of these four detected genes, *fap2* (Mann–Whitney U, $P = 0.008$) was significantly enriched in LS-CRC compared to LS controls, whereas the other three genes were not significantly different ($P > 0.05$) in both LS controls and LS-CRC subjects (**Figure 3.11A**). In comparison, all seven CRC-associated genes were detected in the sporadic CRC cohort, with five genes significantly enriched in CRC subjects, especially those with late-stage CRC subjects (**Figure 3.11B**).

Gene prevalence was similar between the LS and sporadic cohorts (**Figure 3.11C**), indicating a low number of patients with LS as a potential reason for the absence of *fadA2*, *fadA3*, and *FnDps* in the LS cohort. Despite the low sample numbers, *fap2* was highly prevalent in LS CRC and was enriched compared to LS controls. As *fap2* has been reported to play a role in host–microbe adhesion (GalNAc)[142] and immune system modulation (TIGIT)[143], a high abundance of *fap2* may be beneficial for *F. nucleatum* survival and growth in the CRC tumor microenvironment.

Enriched *Fusobacterium* and *fap2* in LS-CRC is in conflict with reports of LS-CRC better prognosis compared to sporadic CRC and worse prognosis for high *Fusobacterium* CRC patients[69,78,88,147]. Notably, our LS cohort had a low detection rate for *fadA* and its homologs, which have been shown to cause DNA damage and CRC cell growth[31,148]. Knockout of *fadA* minimizes *F. nucleatum* effect on CRC progression[148]. Low level of *fadA* and other Fn VF may be associated with better disease prognosis in patients with LS-CRC. Furthermore, a previous study has suggested Fusobacterium and host immunity showed different interaction depending on MSI status, which is one of the hallmarks of LS[149,150]. Thus, a mouse model or human cell line co-culture experiment is needed to elucidate *Fusobacterium* roles in CRC progression and prognosis.

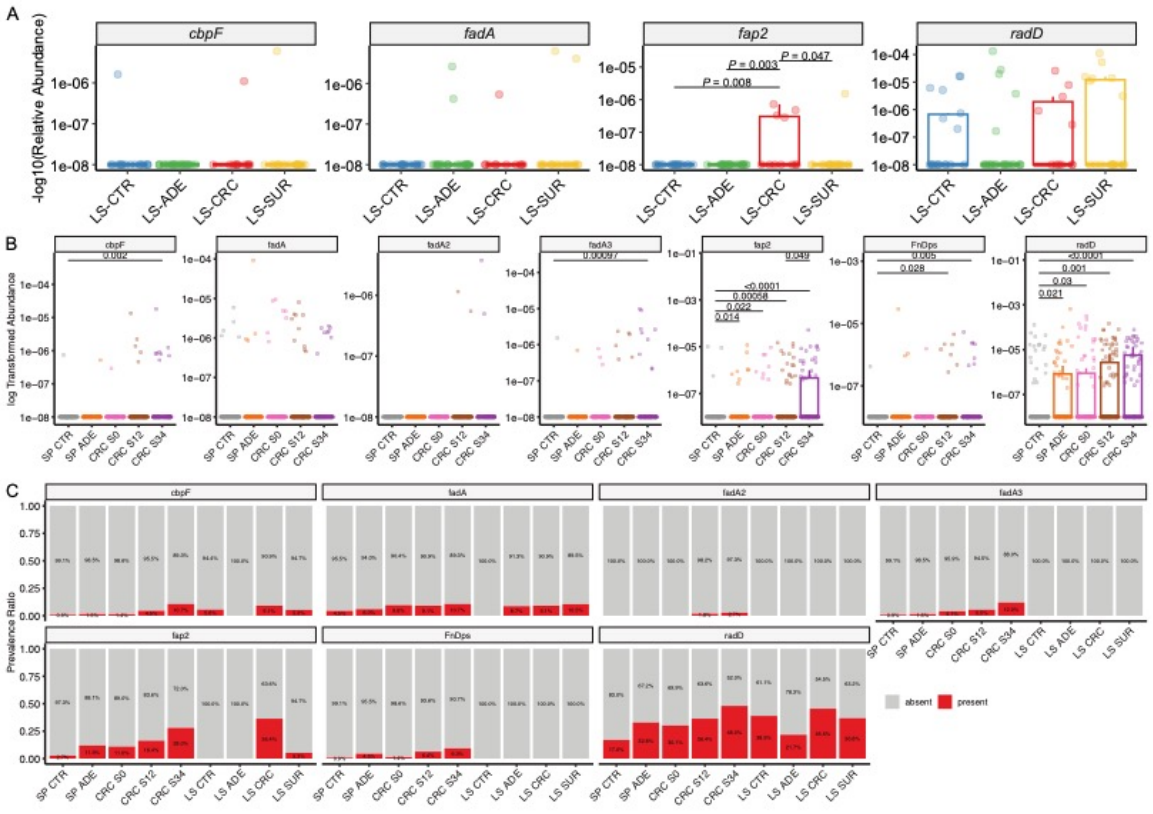


Figure 3.11 Fusobacterium nucleatum virulence factor (Fn VF) abundance and prevalence

A) Boxplot of Fn VFs detected in LS cohort, only comparison (Wilcoxon-Mann-Whitney U test) with statistical significance showed. B) Boxplot of Fn VFs detected in sporadic CRC cohort, comparison with statistical significance showed. C) Fn VFs prevalence in LS and sporadic CRC cohort, red indicates Fn VF detection, gray indicates Fn VF absence.

3.5.2 Escherichia coli and colibactin biosynthesis gene cluster showed no significant change in patients with LS

Other than *Fusobacterium nucleatum* and its adhesins, other microbial genes have also been reported to have a causal relationship with CRC tumorigenesis. Recently, a study reported that colibactin, a genotoxin produced by pks+ *Escherichia coli*, exacerbates dMMR-related mutations[112]. As dMMR is a hallmark of LS, I explored our metagenomic data to identify potential colibactin biosynthesis-related gene enrichment in patients with LS.

Pairwise comparisons of *E. coli* abundance and colibactin biosynthesis-related genes between LS controls and other groups showed no significant changes (**Table 3.1**). A similar analysis of sporadic CRC cohorts also showed no significant changes, except for *clbC* and *clbD,* which showed significant difference in CRC Stage 0 patients and sporadic adenoma patients respectively (**Table 3.2**).

Table 3.1 Pairwise colibactin biosynthesis cluster genes (within LS cohort)

	clbA	*clbB*	*clbC*	*clbD*	*clbF*	*clbG*	*clbH*	*clbI*	*clbJ*	*clbK*	*clbL*	*clbM*	*clbN*	*clbO*	*clbP*	*clbQ*	*clbS*
LS CTR vs LS ADE	0.916	0.459	0.860	0.916	0.684	0.727	0.459	0.727	0.459	0.727	0.429	0.727	0.429	0.727	0.916	0.727	0.403
LS CTR vs LS CRC	0.477	0.477	0.477	0.477	0.477	0.477	0.477	0.477	0.477	0.798	0.477	0.477	0.477	0.477	0.477	0.477	-
LS CTR vs LS SUR	0.330	0.969	0.969	0.330	1.000	0.330	0.969	0.330	0.330	0.330	0.330	0.330	0.330	0.969	0.969	0.330	-

Table 3.2 Pairwise colibactin biosynthesis cluster genes (within non-LS cohort)

	clbA	*clbB*	*clbC*	*clbD*	*clbE*	*clbF*	*clbG*	*clbH*	*clbI*	*clbJ*	*clbK*	*clbL*	*clbM*	*clbN*	*clbO*	*clbP*	*clbQ*	*clbS*
SP CTR vs SP ADE	0.123	0.279	0.691	**0.044**	0.2	0.082	0.199	0.25	0.186	0.342	0.417	0.187	0.129	0.193	0.216	0.104	0.084	0.467
SP CTR vs CRC S0	0.342	0.499	**0.049**	0.505	-	0.502	0.394	0.376	0.778	0.378	0.536	0.557	0.684	0.383	0.51	0.739	0.832	0.402
SP CTR vs CRC S12	0.608	0.871	0.185	0.8	-	0.797	0.55	0.879	0.534	0.984	0.961	0.578	0.942	0.761	0.492	0.704	0.987	0.913
SP CTR vs CRC S34	0.402	0.461	0.202	0.674	-	0.837	0.375	0.103	0.873	0.972	0.842	0.922	0.782	0.998	0.793	0.701	0.316	0.977

3.6 Distinct fecal metabolites and functional gene profile in LS CRC subjects

3.6.1 Higher amino acid degradation and polyamine-related bacterial genes and metabolites in LS-CRC subjects

In addition to direct cell adhesion, gut microbiota-derived metabolites and other metabolic activities may affect the host. To explore the functions of gut microbial genes, I quantified the relative abundance of KEGG Orthologies (KOs). To investigate differences or similarities in other gut microbiota metabolic activities, pairwise comparisons were performed on fecal metabolite concentrations and KO-level functional gene profiles.

A pairwise comparison of the fecal metabolite concentrations between the LS controls and other groups showed 65 metabolites with differential concentrations in at least one comparison (**Appendix 4**). Focusing on LS controls and LS-CRC results, three metabolites (glycerophosphorylcholine, glycerol, and betaine aldehyde) had significantly higher concentrations in LS-CRC. In contrast, 25 metabolites, such as gamma-aminobutyric acid (GABA), malonate, and 2-hydroxyglutarate, had significantly lower concentrations (**Figure 3.12)**. A pairwise comparison of KOs between the LS controls and other groups detected 747 KOs with differential abundance in at least one comparison (**Appendix 5**). Focusing on LS control and LS-CRC comparisons, 271 KOs were significantly more abundant and 63 KOs had a significantly lower abundance in LS-CRC ($P < 0.05$) (**Figure 3.13**).

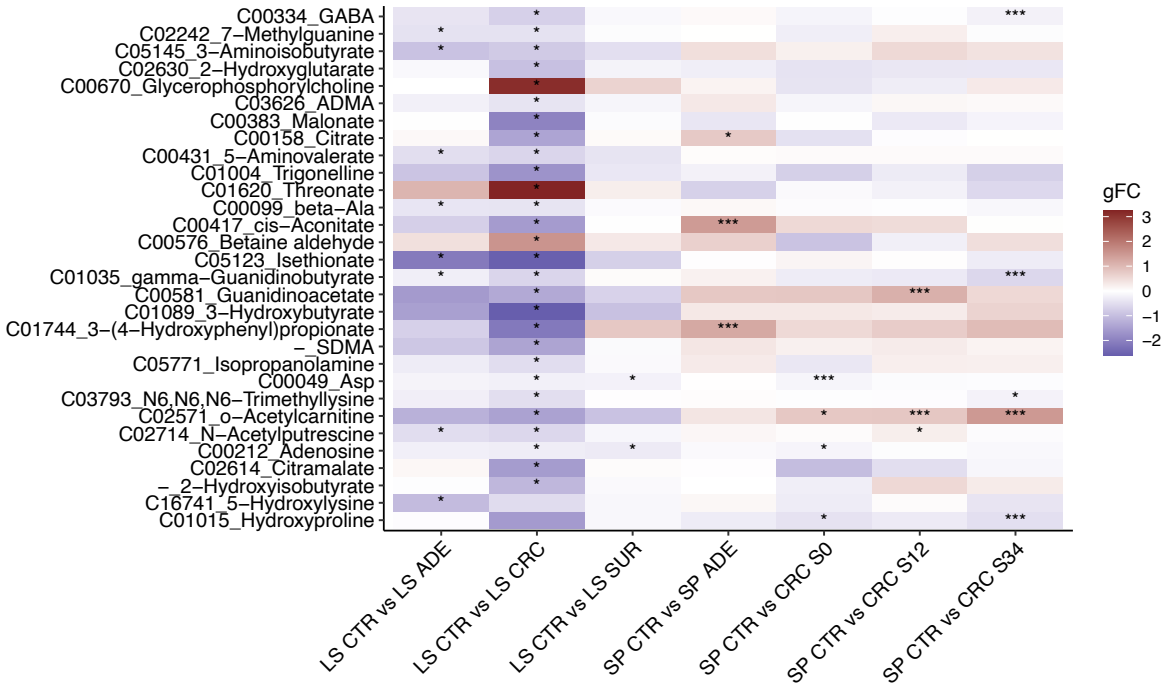


Figure 3.12 Top 30 significantly difference fecal metabolites in LS CRC

Thirty features with the lowest p-value (Wilcoxon-Mann-Whitney U test) in LS control vs LS CRC was picked for visualization. Features were arranged by p-value in ascending order (lowest to highest). Asterisks indicate statistical significance (* $P < 0.05$, *** FDR $P < 0.1$). Color indicates feature abundance differences, estimated by generalized fold-change; red indicate features enriched in case group and blue indicate features enriched in control group. Control vs case was shown in x-axis.

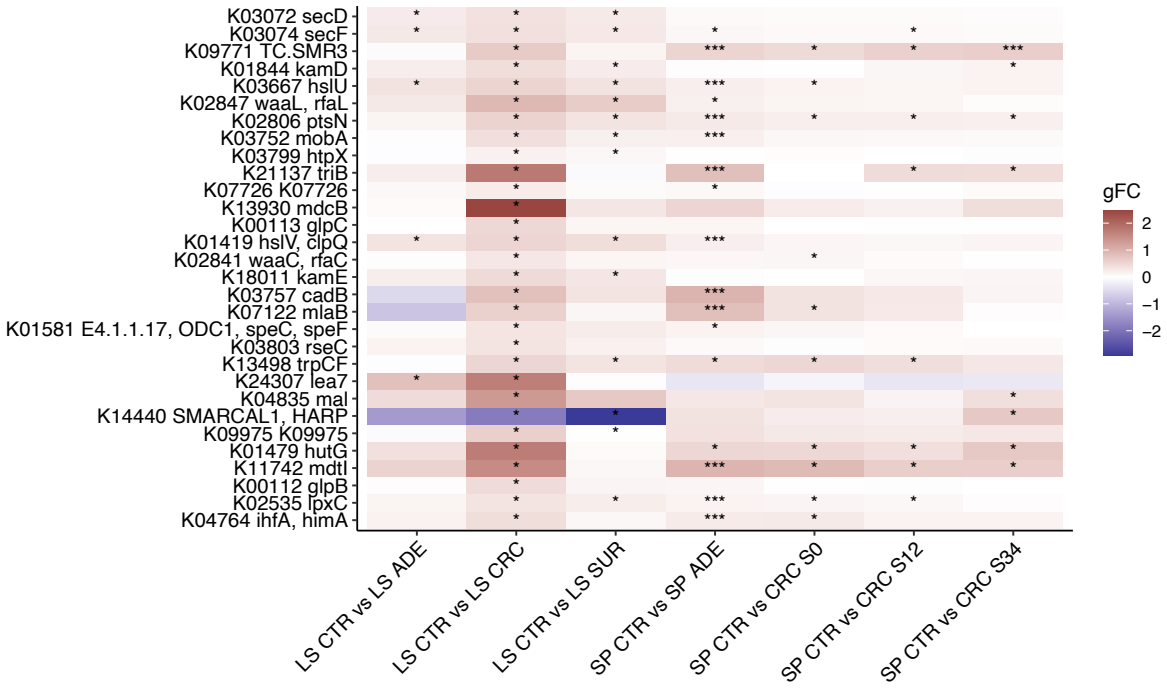


Figure 3.13 Top 30 significantly difference KEGG ORTHOLOGY in LS CRC

Thirty features with the lowest p-value (Wilcoxon-Mann-Whitney U test) in LS control vs LS CRC was picked for visualization. Features were arranged by p-value in ascending order (lowest to highest). Asterisks indicate statistical significance (* $P < 0.05$, *** FDR $P < 0.1$). Color indicates feature abundance differences, estimated by generalized fold-change; red indicate features enriched in case group and blue indicate features enriched in control group. Control vs case was shown in x-axis.

Overrepresentation analysis (ORA) of the KEGG module was performed with fecal metabolites and KOs, showing statistical significance ($P < 0.05$) to determine whether any metabolic function was significantly altered. At KEGG MODULE level, nine modules were significantly over-represented ($q < 0.1$), including “Lysine degradation, L-lysine => succinate” (M00956) and “GABA shunt” (M00027) (**Figure 3.14**). ORA was performed using the EnteroPathway module (EPM) to cover functions undefined in the KEGG MODULE. Three modules (“Lysine degradation, L-lysine => acetoacetyl-coA, butyril-coA” (EPM0634), “Arginine degradation” (EPM0734), and "Ornithine degradation, ornithin => putrescine” (EPM0739)) were significantly overrepresented (**Figure 3.15**).

Overrepresented metabolic modules and components with differential abundance showed that both metabolites and related KOs were altered in LS-CRC (**Figure 3.14, 3.15**). For example, lysine decarboxylase *ldcC* (K01582), which is involved in the first step of the L-lysine degradation pathway, and two other KOs (*gabT* (K07250) and *gabD* (K00135)) that are involved in subsequent reactions in this pathway, were highly abundant in the LS-CRC group. In addition, the fecal concentrations of 5-aminovalerate and 2-hydroxyglutarate, two intermediate metabolites of the L-lysine degradation pathway, were lower in the LS-CRC group. In addition to l-lysine degradation,

gabT (K07250) and *gabD* (K00135) are also involved in GABA shunt formation. Furthermore, fecal GABA concentrations were lower in patients with LS-CRC.

The l-lysine degradation module, which leads to butyrate production, was also overrepresented. The beta-lysine 5,6-aminomutase complex (*kamD* [K01844] and *kamE* [K18011]), L-erythro-3,5-diaminohexanoate dehydrogenase (*kdd* [K18012]), and acetate CoA/acetoacetate CoA-transferase alpha subunit (*atoD* [K01034]) were enriched in LS-CRC.

In addition to L-lysine degradation, ornithine decarboxylase *speC/F* (K01581) and putrescine: ornithine antiporter *potE* (K03756), which are involved in putrescine formation and transport, were also enriched in LS-CRC. In addition, S-adenosylmethionine decarboxylase *speD* (K01611) and spermidine export proteins *mdtI* (K11742) and *mdtJ* (K11743), which are involved in spermidine biosynthesis and export, were enriched in the LS-CRC.

Overall, differential fecal metabolites and functional genes between LS-CRC and LS controls indicated potential shifts in gut microbiota metabolism, specifically higher lysine degradation and polyamine biosynthesis.

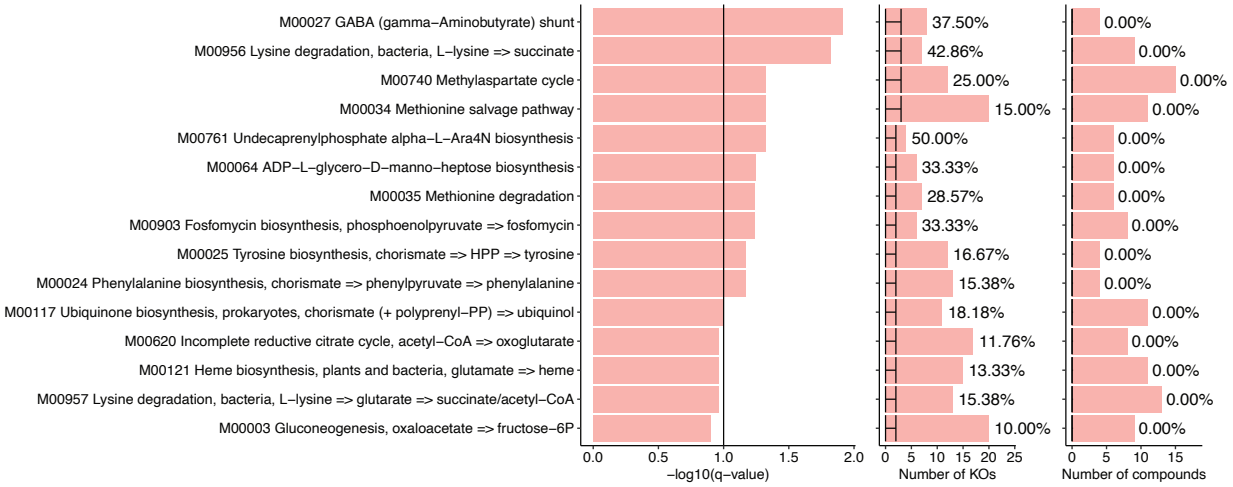


Figure 3.14 KEGG MODULE with over-represented KO and fecal metabolites in LS CRC

Barplots of over-represented KEGG MODULE; number of KOs and metabolite within the module was shown in the x-axis of the middle and left panel, respectively. Percentage indicates the ratio of LS CRC enriched KO or metabolites for each module, e.g. 37.5% of all KO involved in GABA shunt (M00027) was significantly enriched in LS CRC, with no enriched metabolites (GABA was significantly depleted in LS CRC, **Figure 3.12**).

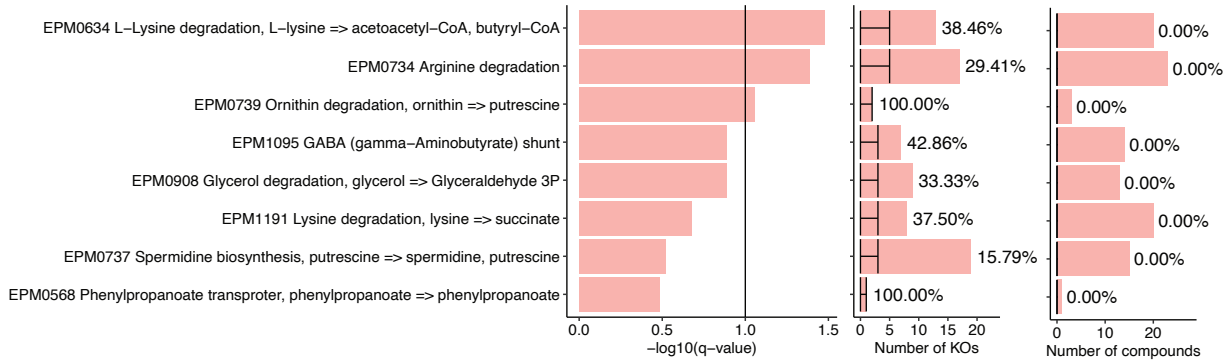


Figure 3.15 EnteroPathway module (EPM) with over-represented KO and fecal metabolites in LS CRC

Barplots of over-represented EPM; number of KOs and metabolite within the module was shown in the x-axis of the middle and left panel, respectively. Percentage indicates the ratio of LS CRC enriched KO or metabolites for each module.

3.6.2 Distinct carbohydrate utilization of LS subject's gut microbiota

Previous studies have reported CRC-associated dysbiosis resulted in enriched mucolytic microbiota and depleted fiber-degrading microbiota. Some fiber-degrading microbiota ferments fiber to short-chain fatty acid (SCFA), which have been associated with CRC, although the exact roles is still unclear[36,37,39]. To identify potential disturbance in the gut microbiota carbohydrate utilization, differential abundance analysis on carbohydrate active enzymes (CAZy) profile was performed.

CAZy profile differential abundance analysis results showed 20 CAZy with significant difference in LS CRC subjects compared to LS controls (**Figure 3.16, Appendix 6**). CAZy enriched in LS CRC subjects included those involved in pectin and peptidoglycan metabolism. In contrast, CAZy involved in host glycan degradation was depleted in LS CRC subjects. Of interest, CAZy involved in mucin degradation, such as GH20, GH33 and CBM32[118], showed significant enrichment in sporadic CRC, but not in LS CRC.

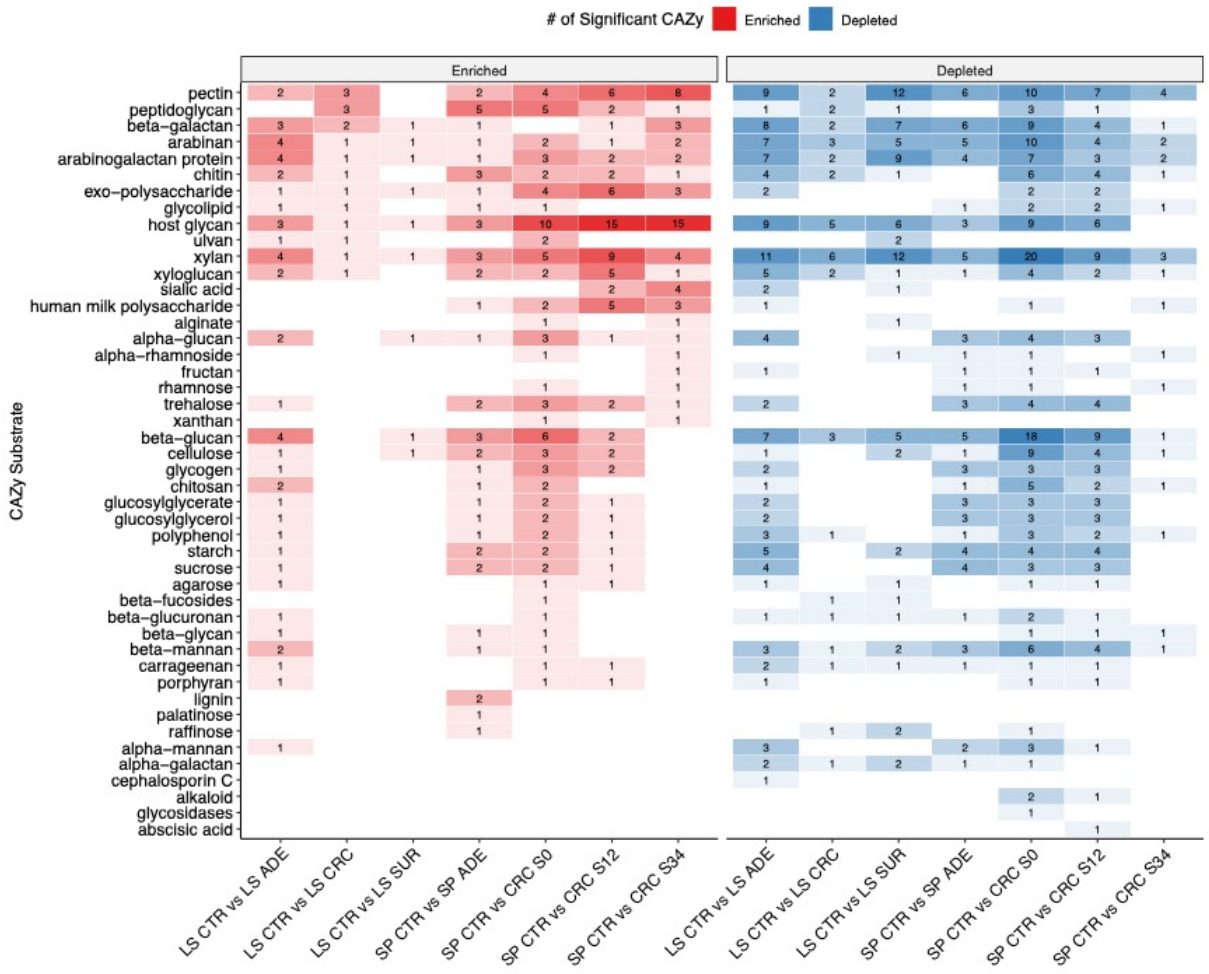


Figure 3.16 Number of significant CAZy for each substrate

Summary of significant CAZy for each putative substrate, splitted into enriched (right) and depleted (left). Number in heatmap tiles represent number of significant CAZy.

3.6.3 Distinct secreted proteins of LS subject's gut microbiota

Based on KO profile comparison and CAZy profile comparison results, LS subject's gut microbiota showed functional shifts between control and CRC subjects. As secreted or membrane proteins have been reported to play important roles in host-microbe interaction, potential secreted proteins were extracted and compared to discover potential associations with CRC progression in the LS cohort.

From differential abundance analysis results, several features showed significant enrichment in LS CRC subjects compared to LS controls, mostly consisting of transporters and membrane proteins (**Figure 3.17, Appendix 7**). Features enriched in LS CRC showed similar enrichment trends in sporadic adenoma subjects, indicating changes might occur prior to CRC formation. In contrast,

the resemblance in enrichment trends were less prevalent when compared to late-stage sporadic CRC.

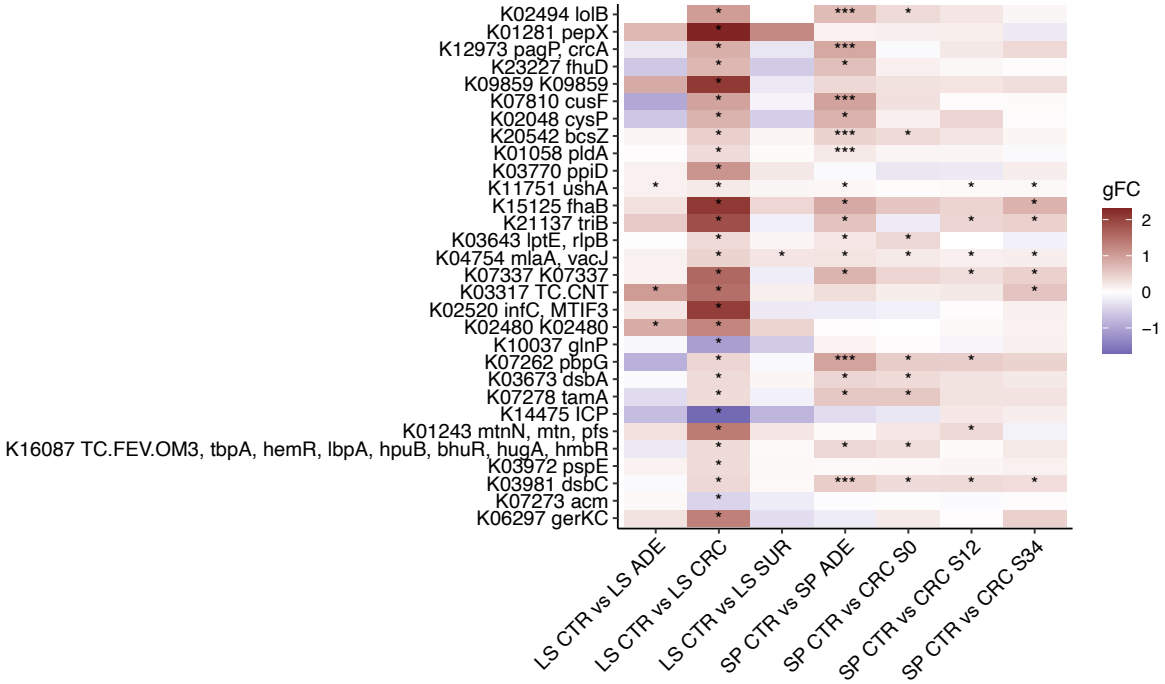


Figure 3.17 Top 30 significantly different putative secreted/trans-membrane proteins in LS CRC

Thirty features with the lowest p-value (Wilcoxon-Mann-Whitney U test) in LS control vs LS CRC was picked for visualization. Features were arranged by p-value in ascending order (lowest to highest). Asterisks indicate statistical significance (* $P < 0.05$, *** FDR $P < 0.1$). Color indicates feature abundance differences, estimated by generalized fold-change; red indicate features enriched in case group and blue indicate features enriched in control group. Control vs case was shown in x-axis.

Over-represented analysis at KEGG MODULE levels of secreted or membrane proteins and fecal metabolites showed 8 significantly enriched KEGG MODULEs (**Figure 3.18**), 2 of which were related to sulfur metabolism (M00616 Sulfate-sulfur assimilation and M00609 Cysteine biosynthesis). At EnteroPathway MODULE (EPM) level, only EPM1051 Phosphatidylcholine degradation showed statistical significance (**Figure 3.19**).

Finally, some features showing differential abundance in LS control vs LS CRC comparison did not belong to a KEGG MODULE or EPM, thus over-represented analysis was not capable of detecting potential enriched transporter proteins group.

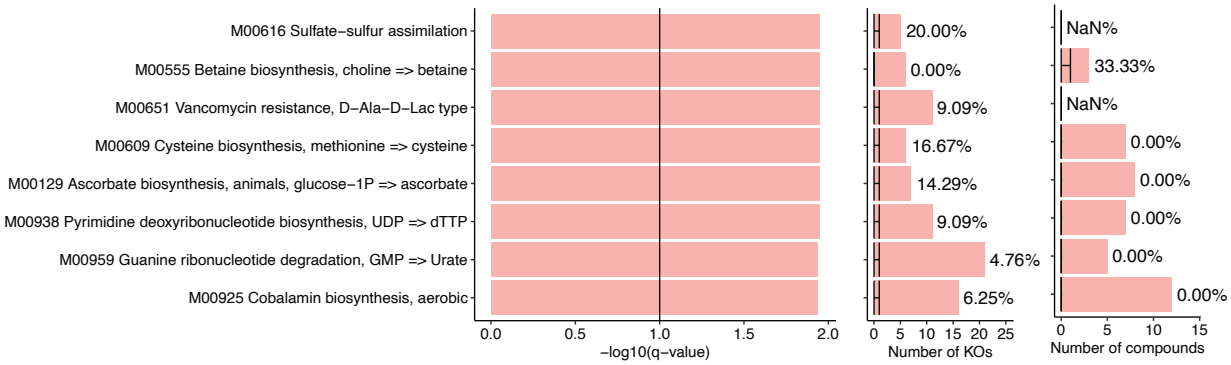


Figure 3.18 KEGG MODULE with over-represented secreted proteins and fecal metabolites in LS CRC

Barplots of over-represented KEGG MODULE; number of putative secreted or trans-membrane KOs and metabolite within the module was shown in the x-axis of the middle and left panel, respectively. Percentage indicates the ratio of LS CRC enriched KO or metabolites for each module.

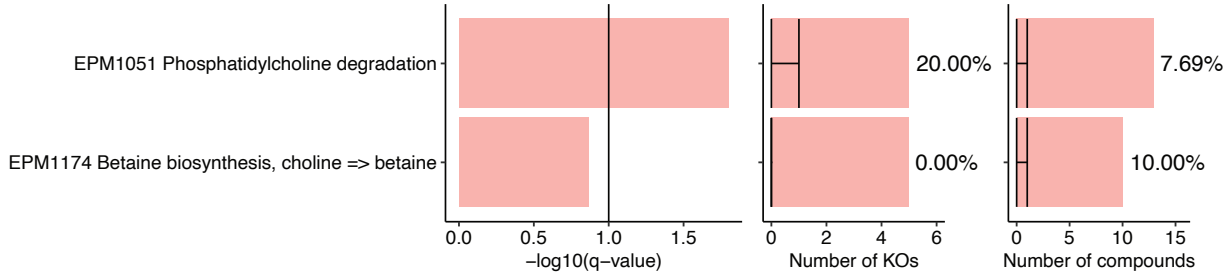


Figure 3.19 EPM with over-represented secreted proteins and fecal metabolites in LS CRC

Barplots of over-represented EPM; number of putative secreted or trans-membrane KOs and metabolite within the module was shown in the x-axis of the middle and left panel, respectively. Percentage indicates the ratio of LS CRC enriched KO or metabolites for each module.

3.7 Minimal gut microbiota dysbiosis in non-tumor LS control subjects compared to age and sex-matched non-LS healthy controls

3.7.1 In silico cohort matching to consider potential confounding factors

LS controls (n = 18) and non-LS controls (n = 112) was compared to examine the gut microbiota and fecal metabolite changes that occurred prior to adenoma formation in patients with LS. To account for sample number imbalance and confounding factors (**Figure 3.20B top**), non-LS control samples were matched to LS control samples based on age and sex distributions, cutting down the number of non-LS controls to 18 subjects (**Figure 3.20B bottom**). Cohort-matched non-LS and LS no longer showed significant age difference, although other variables, such as chicken meat consumption still showed statistical significance (**Figure 3.20A**)

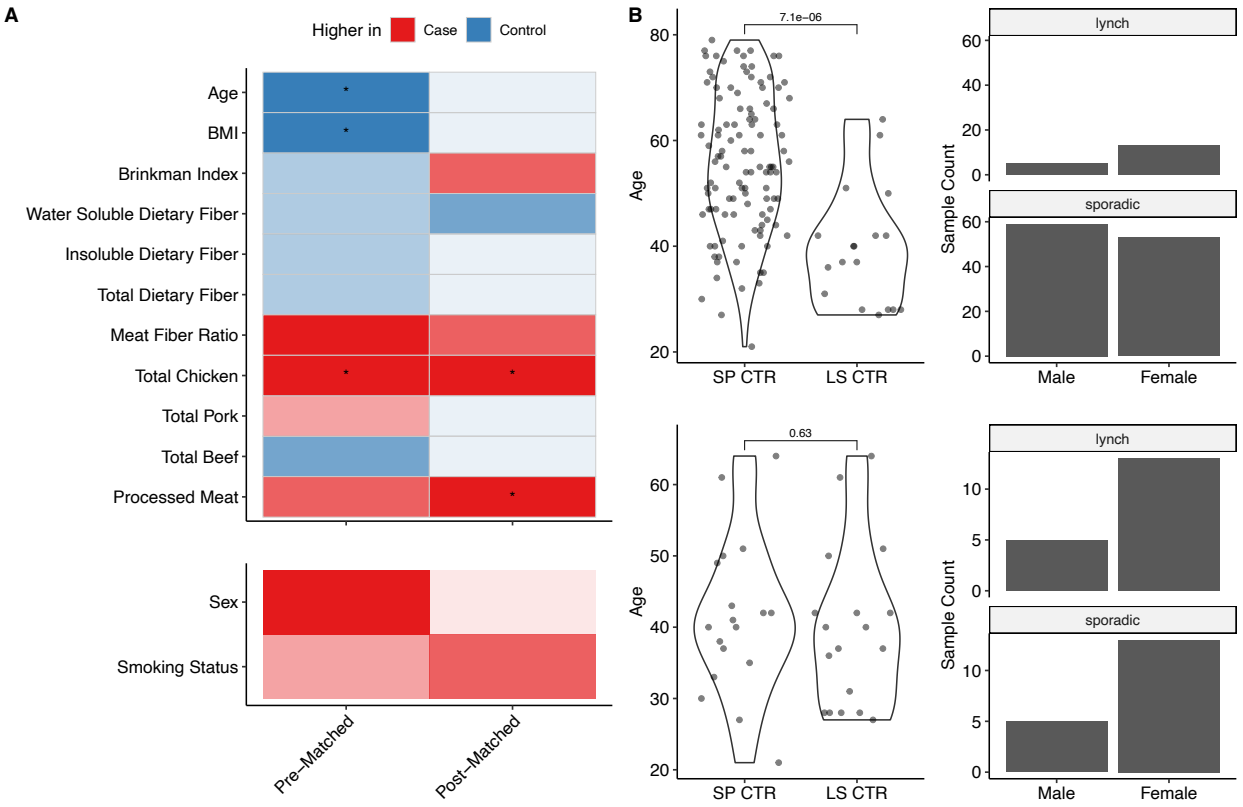


Figure 3.20 Metadata analysis between non-LS and LS controls, pre- and post- cohort matching

A) Covariate comparison between non-LS and LS control before and after cohort-matching; asterisk represents statistical significance ($P < 0.05$). B) Age and sex distribution before (top) and after cohort matching (bottom).

3.7.2 No significant enrichment of CRC-associated gut microbiota detected in LS controls

A comparison between the LS and non-LS controls showed 24 species with differential abundances (**Figure 3.21, Appendix 8**). Of the 24 species, three (*Bacteroides helcogenes*, *Bacteroides plebeius*, and *Cronobacter turicensis*) were enriched in LS controls. Differential abundance of *Fusobacterium* spp. and other CRC-related gut microbes in LS controls was not observed, indicating differential abundance detected between LS control and LS CRC was not caused by LS-specific baseline gut microbiota abundance (**Figure 3.21**).

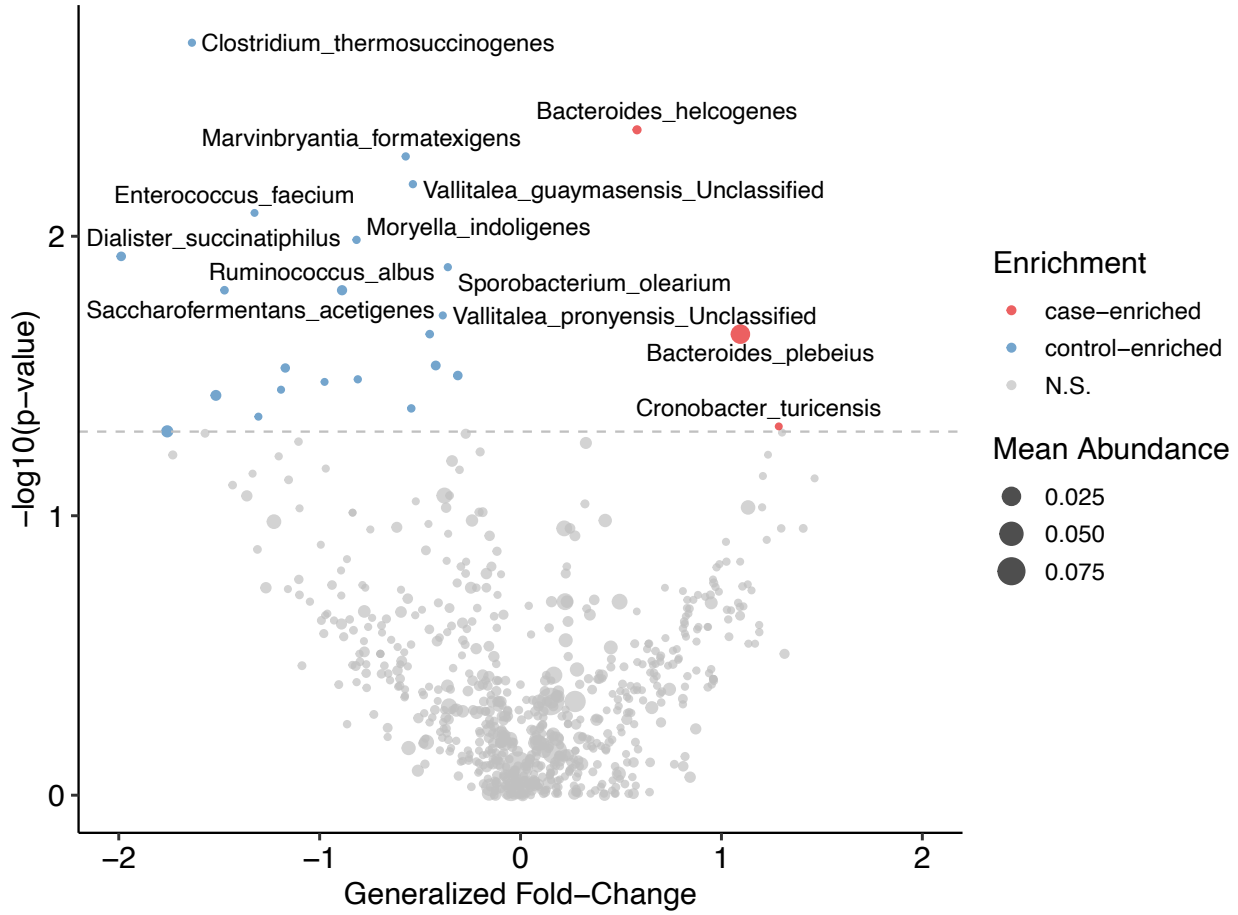


Figure 3.21 Cohort matched non-LS vs LS controls LTP species comparison

Volcano plot of species-level profile comparison between non-LS and LS controls. Point size represents feature's relative abundance, color represents enrichment status; red = enriched in LS control, blue = enriched in non-LS control; gray = no significant difference detected. The x-axis represents abundance difference estimated by generalized fold-change (positive = higher in LS control; negative = higher in non-LS control). The y-axis represents statistical significance estimated by -$\log_{10}$(*P*-value).

3.7.3 Distinct fecal metabolite profile, but minimal functional gene difference was detected in LS controls

A comparison of fecal metabolite concentrations showed that 57 metabolites had significantly different concentrations between sporadic and LS controls, 52 of which were higher in the LS controls (**Figure 3.22, Appendix 9**). Metabolites with higher concentrations in LS controls included several known oncometabolites, such as asymmetric dimethylarginine (ADMA) and spermidine[151–154]. Differential abundance analysis of the KO profile showed 93 KOs higher in LS controls and 210 KOs lower in LS controls than in non-LS controls (**Figure 3.23, Appendix 10**).

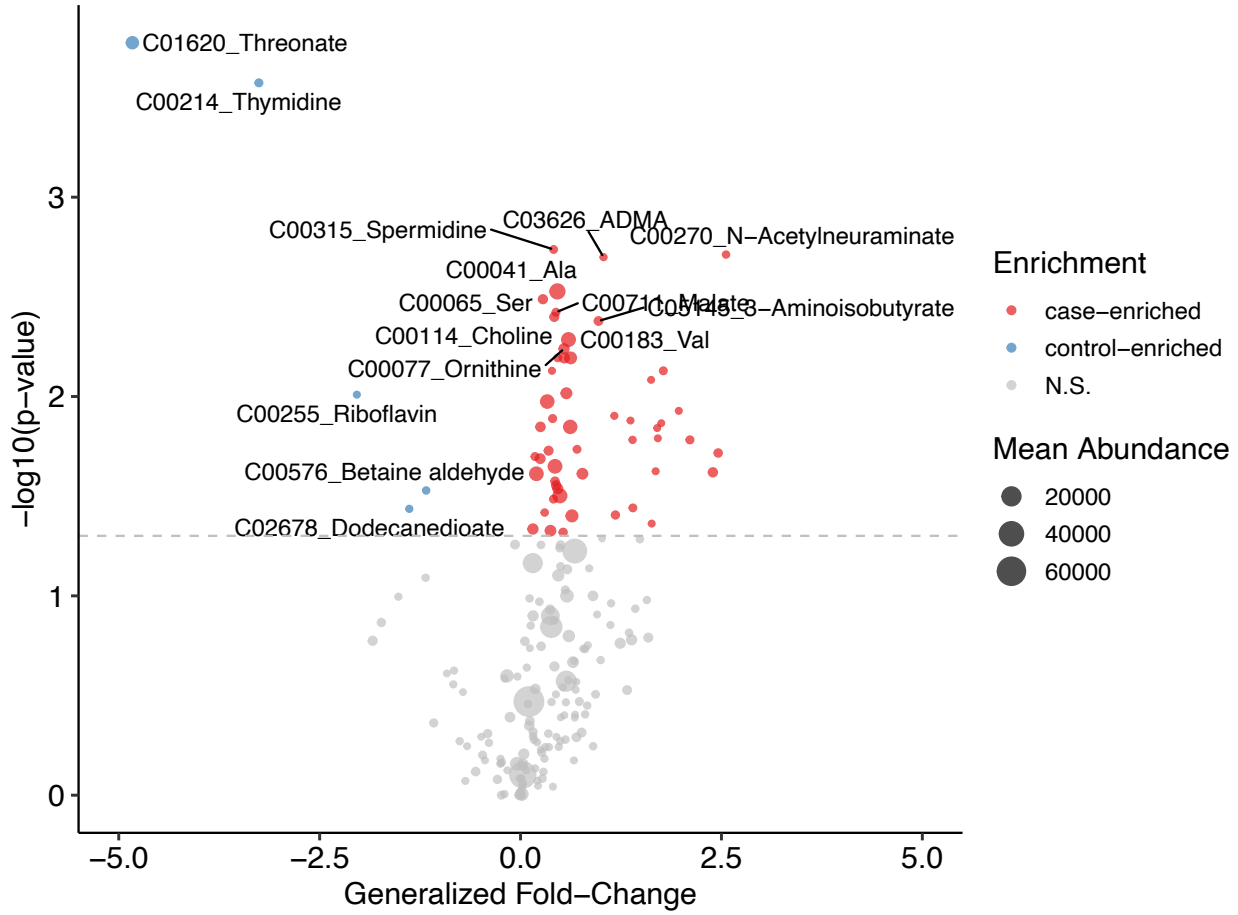


Figure 3.22 Cohort matched non-LS vs LS controls fecal metabolite comparison

Volcano plot of fecal metabolite profile comparison between non-LS and LS controls. Point size represents feature's relative abundance, color represents enrichment status; red = enriched in LS control, blue = enriched in non-LS control; gray = no significant difference detected. The x-axis represents abundance difference estimated by generalized fold-change (positive = higher in LS control; negative = higher in non-LS control). The y-axis represents statistical significance estimated by -$\log_{10}$(*P*-value).

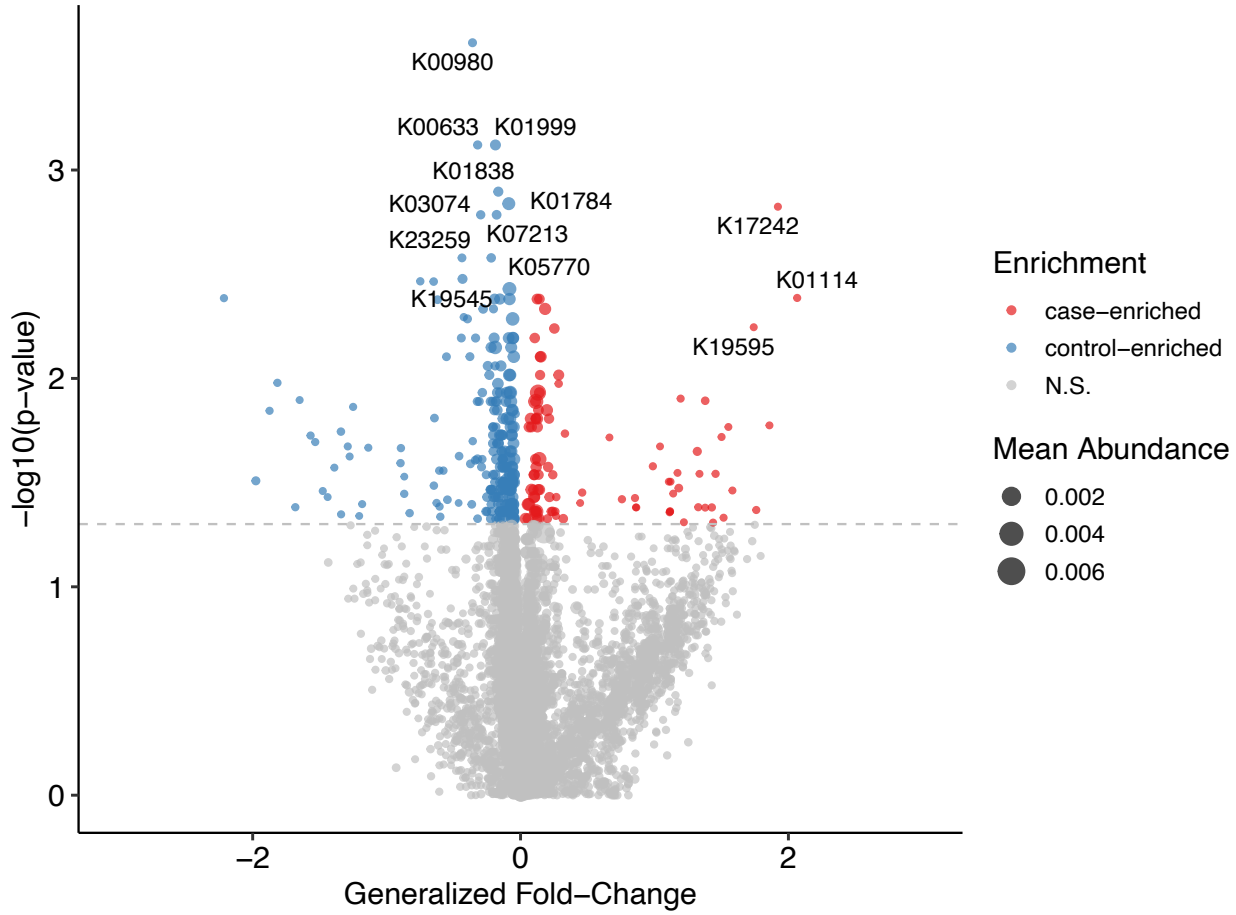


Figure 3.23 Cohort matched non-LS vs LS controls KEGG ORTHOLOGY comparison

Volcano plot of KO profile comparison between non-LS and LS controls. Point size represents feature's relative abundance, color represents enrichment status; red = enriched in LS control, blue = enriched in non-LS control; gray = no significant difference detected. The x-axis represents abundance difference estimated by generalized fold-change (positive = higher in LS control; negative = higher in non-LS control). The y-axis represents statistical significance estimated by -$\log_{10}$(*P*-value).

ORA on fecal metabolites and KOs showed 25 KEGG MODULES were significantly over-represented, including "Arginine biosynthesis, glutamate => arginine" (M00845), "Arginine biosynthesis, ornithine => arginine" (M00844), "Ornithine biosynthesis, glutamate => ornithine" (M00028), and "Polyamine biosynthesis, arginine => spermidine" (M00133) (**Figure 3.24**). Three modules in EnteroPathway were significantly overrepresented ("Arginine biosynthesis, ornithine => arginine" (EPM1197), "Conjugated CA deconjugation" (EPM0129), and "GNB biosynthesis" (EPM0819)) (**Figure 3.25**). However, the ORA results were mostly driven by shifts in fecal metabolite concentrations with minimal changes in functional gene abundance (**Figure 3.24, 3.25**).

Differential abundance analysis and over-represented analysis showed fecal concentrations of amino acids (e.g., ornithine) and polyamines (e.g., spermidine and agmatine) were higher in LS controls than in non-LS controls. However, most microbial genes involved in amino acid or

polyamine metabolism showed no significant changes in LS controls. Similarly, fecal cholate, taurine, and glycine were enriched in the LS controls, with no significant enrichment detected for bile acid hydrolases. Notably, deoxycholate and the corresponding *bai* operon-related genes were not significantly enriched in LS controls. Bile acid metabolism has been associated with increased CRC risk, especially DCA-induced DNA oxidative damage[155]. As DNA oxidative damage is one of the repair targets of mismatch repair genes[156], the accumulation of bile acids might significantly impact CRC tumorigenesis in patients with LS.

Overall, a comparison between LS and non-LS controls showed minimal changes in gut microbiota composition and function. However, distinct fecal metabolite profiles and gut microbiota activities were observed, indicating that a distinct LS-specific intestinal environment may be present prior to adenoma formation. Distinct immune profiles have been reported in LS patients with dMMR colonic crypts, even prior to adenoma and carcinoma formation[138]. As activated immune systems have been reported to induce metabolic reprogramming[157], the observed shifts in fecal metabolites might be a marker of the activated host immune system triggered by early dMMR mutations.

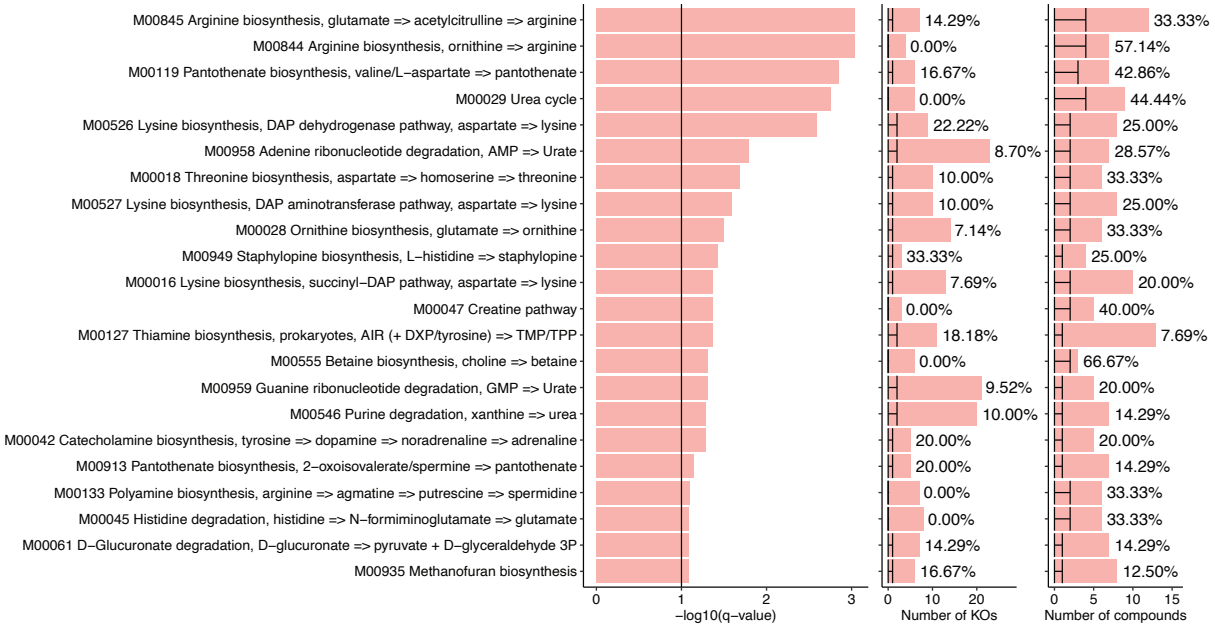


Figure 3.24 KEGG MODULE with over-represented KO and fecal metabolites in LS controls

Barplots of over-represented KEGG MODULE; number of KOs and metabolite within the module was shown in the x-axis of the middle and left panel, respectively. Percentage indicates the ratio of LS CRC enriched KO or metabolites for each module.

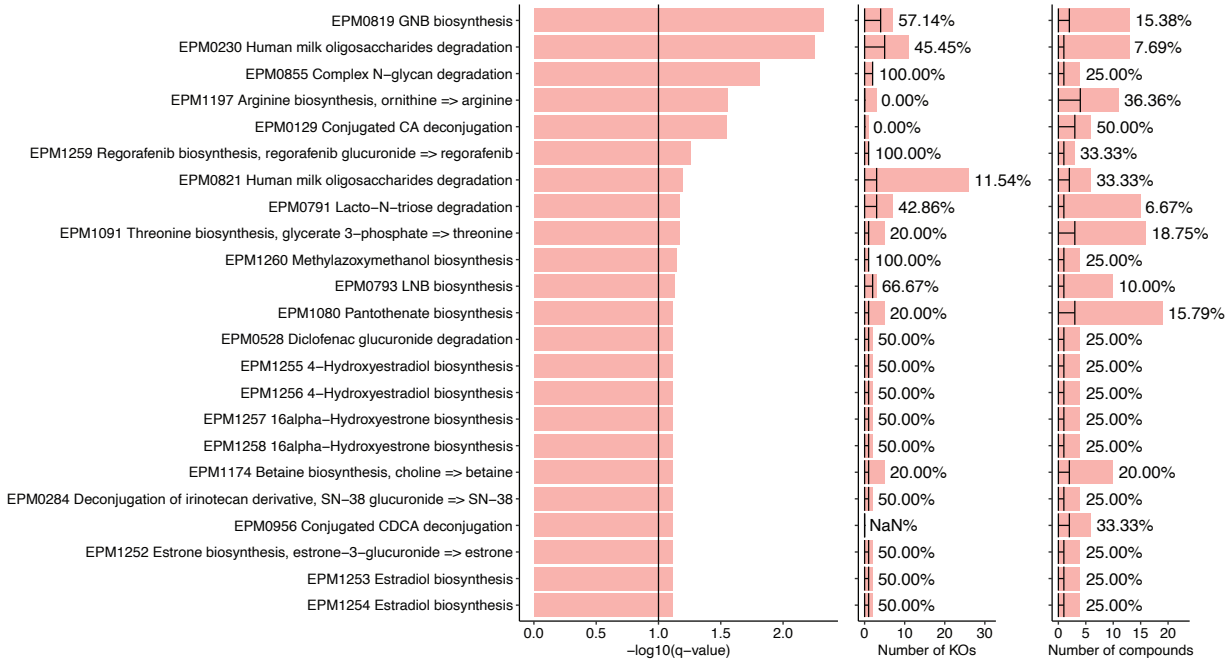


Figure 3.25 EPM with over-represented KO and fecal metabolites in LS controls

Barplots of over-represented EPM; number of KOs and metabolite within the module was shown in the x-axis of the middle and left panel, respectively. Percentage indicates the ratio of LS CRC enriched KO or metabolites for each module.

3.7.4 Enriched host glycan related CAZy in LS controls

N-acetylneuraminate, a known mucin substrate, was identified to be enriched in LS controls (**Figure 3.22**) and EPM-level over-representation analysis indicate increased host glycan utilization by the gut microbiota (**Figure 3.25**). Therefore, I explored CAZy annotated gene profile to identify potential carbohydrate utilization pattern distinct to LS subjects even prior to adenoma formation.

CAZy profile comparison results showed LS controls has a distinct CAZy profile compared to non-LS controls (**Figure 3.26, Appendix 11**). Of interest, CAZy associated with host glycan degradation was enriched in LS controls, indicating increased host glycan utilization by the LS control subject's gut microbiota even prior to adenoma formation (**Figure 3.25**). Enriched host glycan utilization might reduce intestinal barrier, thus increasing the rate of host-microbe interaction, cascading into chronic intestinal inflammation. Chronic intestinal inflammation has been associated with CRC and there are reports of colitis-induced CRC and gut microbiota association.

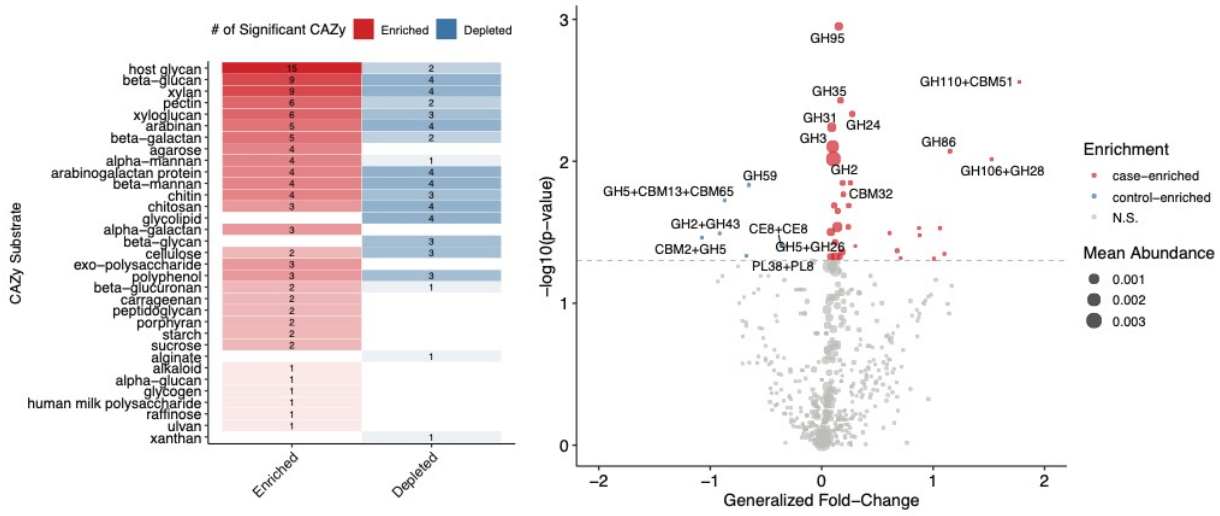


Figure 3.26 Non-LS vs LS controls CAZy comparison and putative CAZy substrate

A) Number significant CAZy for each putative substrate, splitted into LS control enriched (right) and LS control depleted (left). Number in heatmap tiles represent number of significant CAZy. B) Volcano plot of KO profile comparison between non-LS and LS controls. Point size represents feature's relative abundance, color represents enrichment status; red = enriched in LS control, blue = enriched in non-LS control; gray = no significant difference detected. The x-axis represents abundance difference estimated by generalized fold-change (positive = higher in LS control; negative = higher in non-LS control). The y-axis represents statistical significance estimated by -$\log_{10}$(*P*-value).

Chapter 4

Dissertation Summary and Further Direction

In this study, fecal samples were collected from 71 LS carriers of varying CRC progression status. The LS cohort included patients without adenoma or carcinoma and no prior history of adenoma or carcinoma formation, those with colorectal adenoma, those with CRC, and those with a history of colectomy due to past carcinoma formation. Furthermore, sporadic CRC cohort with identical sampling protocol is utilized as a comparison cohort to identify potential LS-specific gut microbiota association. To my knowledge, this study is the first to analyze the changes in gut microbiota and fecal metabolite across all stages of adenoma-carcinoma progression in LS, along with a comparison with non-LS CRC. Based on this study's results and known findings, I hypothesize that in the context of CRC progression in LS subjects, host genetic is the main driving factor, with host-microbe interaction potentially affecting subsequent disease progression or treatment response (**Figure 4.1**).

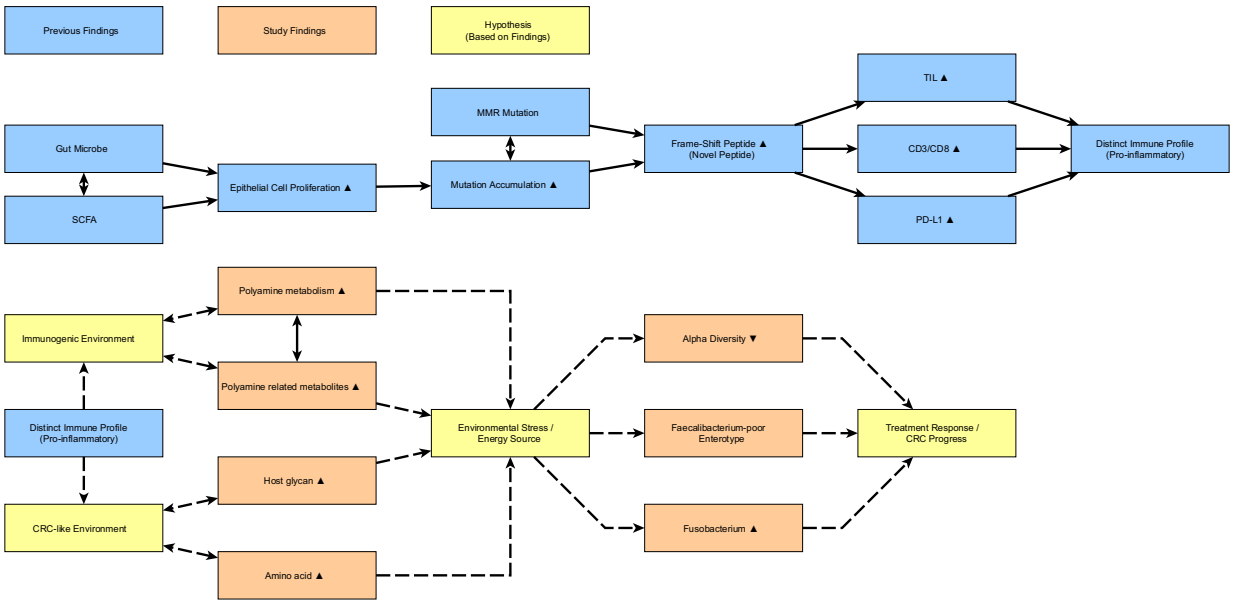


Figure 4.1 Hypothetical host-microbe interaction in the context of LS CRC progression

Schematic hypothesis of host-microbe interaction in CRC progression of LS subjects. Colors indicates statement sources, such as previous study (blue), observations from this study (orange), and observation-based hypothesis (yellow). Solid lines represent known associations and dashed lines represent putative associations. Triangles represent comparison results between LS and non-LS subjects, with upright triangle representing higher level in LS and upside down triangle representing lower level in LS.

From compositional data analysis, LS cohort was shown to have a distinct gut microbiota compared to the sporadic CRC cohort. Specifically, LS cohort has lower alpha-diversity and higher proportion of *Faecalibacterium*-poor enterotype, which were both associated with intestinal inflammation. Beta-diversity analysis supported this observation with LS status detected as

capable of explaining the gut microbiota composition variance. However, beta-diversity analysis on non-LS controls and LS controls showed LS status was not explanatory of gut microbiota, indicating CRC progression confounds earlier observation. Additionally, within LS cohort, no covariates were explanatory of the gut microbiota composition variance, with MMR mutation as the most explanatory covariate. As MMR mutation has been associated with CRC risk in LS subjects, there is a need to deconfound MMR mutation and CRC status for a more robust association with gut microbiota.

One of the most important findings of this study is the robust signal of *Fusobacterium* spp. enrichment in LS CRC subjects. Differential abundance analysis, trend test, association test and ML-based feature screening results suggest that *Fusobacterium* spp., including *F. nucleatum* was enriched in LS CRC, but not in LS adenoma and LS colectomy, in line with previous LS gut microbiome studies. Furthermore, no other CRC-associated gut microbiota was detected to be enriched in LS CRC, suggesting LS CRC intestinal environment might be advantageous to *Fusobacterium* spp. survival. Nevertheless, as no longitudinal data was used in this study, I am not sure whether this enrichment occurs before or after carcinoma formation.

Although *Fusobacterium* spp. was enriched in LS CRC, *F. nucleatum* virulence factors (Fn VFs) had low prevalence in this study cohort. Of seven Fn VFs explored in this study, only *fap2*, an adhesin associated with CRC cell binding and immune system evasion, was detected to be enriched in LS CRC. In comparison, other Fn VFs, such as *fadA*, *radD*, *cbpF*, and *Fn-Dps*, were significantly enriched in sporadic CRC cohort. As Fn VFs detection rate was similar in LS and sporadic CRC cohort, the limited number of LS samples might not have enough detection power for Fn VFs comparison.

F. nucleatum has been associated with cancer metastasis, chemoresistance, and immune checkpoint inhibitor (ICI) response. Chemoresistance and ICI response have been associated with CRC, especially in patients with high microsatellite instability (MSI-H). Furthermore, a higher *F. nucleatum* load has been reported in MSI-H CRC. As MSI-H is a hallmark of LS-CRC and our observations showed a high abundance of *F. nucleatum* in LS CRC, LS-CRC might be utilized as a model case for MSI-H CRC progression and treatment response.

Differential abundance analysis of fecal metabolites and gene profiles revealed putative functional shifts in LS CRC gut microbes compared to LS controls, such as overrepresentation of the lysine degradation pathway, GABA shunts, and polyamine biosynthesis. Higher amino acid utilization as an energy source and increased polyamine metabolism by gut microbes may be adaptive responses to CRC-driven changes in the intestinal environment.

Notably, CAZy related to host glycan was not enriched in LS CRC, which were unexpected as MSI-H CRC is characterized as mucinous tumor. In comparison, host glycan related CAZy was

enriched in late-stage sporadic CRC. However, CAZy comparison between non-LS and LS controls showed enriched host glycan related CAZy in LS controls, indicating high mucin utilization of LS gut microbiota might occurs earlier than expected.

Another important finding of this study was the discovery of distinct fecal metabolite compositions in patients with LS without cancerous lesions. Fecal concentrations of amino acids (e.g., ornithine), polyamines (e.g., spermidine and agmatine), bile acid related metabolite (taurine, glycine, cholate) were higher in LS controls than in non-LS controls. However, most microbial genes involved in amino acid or polyamine metabolism showed no significant changes in LS controls. In addition, no significant differences of CRC-associated microbiota were detected in non-LS vs LS controls comparison.

Based on these observations, host genetics is likely the main driver of CRC tumorigenesis in LS, and shifts in gut microbiota composition and function occur in response to changes in the intestinal environment caused by intestinal inflammation or tumor formation. Specific members of the gut microbiota may have specific genes or metabolic pathways that allow them to thrive in environments, such as *fap2*, L-lysine degradation, GABA shunts, and polyamine biosynthesis. Surviving gut microbiota may play a role in CRC progression, metastasis, disease prognosis, and treatment response.

As this study only utilized cross-sectional data of limited size, it is not possible to study the causality of my findings. Thus, a longitudinal data from a larger cohort is imperative to validate this study's findings. For a more reductionist approach, a co-culture experiment of human intestinal cell or tissue and *F. nucleatum* is ideal to study the potential interaction and effect on CRC progression or response to treatment. Specifically, it would be interesting to study whether *fadA* or *fap2* mutant *F. nucleatum* strains would affect CRC progression and treatment response. Furthermore, to test on the hypothesis of immunogenic CRC environment effect on gut microbiota, it would be interesting to perform co-culture experiment of human CRC cell and common gut microbiota strains. However, as cross-feeding and microbe-microbe interaction plays a big role in gut microbiome composition and function, it is hard to perform this co-culture experiment with current frameworks.

Other limitation of this study is the incapability to consider all potential confounding factors, as patients were recruited naturally instead of actively recruiting specific demographic. Thus, it is of great interest to recruit a specific cohort to account for confounding factors, such as MMR mutation variants, diet, age, sex, and socioeconomic factors. Furthermore, as the study only utilized Japanese subjects, it would be interesting to see the effect of human genetics by comparing with a more genetically varied cohort.

Reference

1. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nature Medicine* **24**, 392–400 (2018).
2. Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *New England Journal of Medicine* **375**, 2369–2379 (2016).
3. Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 103 (2020).
4. Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
5. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
6. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
7. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679 (2019).
8. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
9. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine* **9**, (2017).
10. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

11. Vieira-Silva, S. *et al.* Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat Microbiol* **4**, 1826–1831 (2019).

12. Li, Q. *et al.* Streptococcus thermophilus Inhibits Colorectal Tumorigenesis Through Secreting β-Galactosidase. *Gastroenterology* **160**, 1179-1193.e14 (2021).

13. Long, X. *et al.* Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat Microbiol* **4**, 2319–2330 (2019).

14. Irrazabal, T. *et al.* Limiting oxidative DNA damage reduces microbe-induced colitis-associated colorectal cancer. *Nature Communications* **11**, 1802 (2020).

15. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**, 766 (2014).

16. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

17. Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).

18. Lee, K. A. *et al.* Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat Med* **28**, 535–544 (2022).

19. Sorbara, M. T. & Pamer, E. G. Microbiome-based therapeutics. *Nat Rev Microbiol* **20**, 365–380 (2022).

20. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* **2**, 3 (2012).

21. MetaHIT Consortium *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* **32**, 822–828 (2014).

22. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).

23. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**, 833–844 (2017).

24. Aguiar-Pulido, V. *et al.* Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data. *Evol Bioinform Online* **12s1**, EBO.S36436 (2016).

25. Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* **18**, 747–756 (2021).

26. Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol Case Stud* **1**, a000588 (2015).

27. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).

28. Fearon, E. R. Molecular Genetics of Colorectal Cancer. *Annual Review of Pathology: Mechanisms of Disease* **6**, 479–507 (2011).

29. Nguyen, L. H., Goel, A. & Chung, D. C. Pathways of Colorectal Carcinogenesis. *Gastroenterology* **158**, 291–302 (2020).

30. Yu, T. *et al.* Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell* **170**, 548-563.e16 (2017).

31. Rubinstein, M. R. *et al.* Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. *Cell Host & Microbe* **14**, 195–206 (2013).

32. Ternes, D. *et al.* Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends in Microbiology* **28**, 401–423 (2020).

33. Chung, L. *et al.* Bacteroides fragilis Toxin Coordinates a Pro-carcinogenic Inflammatory Cascade via Targeting of Colonic Epithelial Cells. *Cell Host & Microbe* **23**, 203-214.e5 (2018).

34. Dalmasso, G., Cougnoux, A., Delmas, J., Darfeuille-Michaud, A. & Bonnet, R. The bacterial genotoxin colibactin promotes colon tumor growth by modifying the tumor microenvironment. *Gut Microbes* **5**, 675–680 (2014).

35. Brennan, C. A. & Garrett, W. S. Fusobacterium nucleatum — symbiont, opportunist and oncobacterium. *Nat Rev Microbiol* **17**, 156–166 (2019).

36. Bultman, S. J. & Jobin, C. Microbial-Derived Butyrate: An Oncometabolite or Tumor-Suppressive Metabolite? *Cell Host & Microbe* **16**, 143–145 (2014).

37. Okumura, S. *et al.* Gut bacteria identified in colorectal cancer patients promote tumourigenesis via butyrate secretion. *Nat Commun* **12**, 5674 (2021).

38. Wu, X. *et al.* Effects of the intestinal microbial metabolite butyrate on the development of colorectal cancer. *Journal of Cancer* **9**, 2510–2517 (2018).

39. Belcheva, A. *et al.* Gut Microbial Metabolism Drives Transformation of Msh2-Deficient Colon Epithelial Cells. *Cell* **158**, 288–299 (2014).

40. Yang, J. *et al.* High soluble fiber promotes colorectal tumorigenesis through modulating gut microbiota and metabolites in mice. *Gastroenterology* S0016508523051399 (2023) doi:10.1053/j.gastro.2023.10.012.

41. Nakatsu, G. *et al.* Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nature Communications* **6**, (2015).

42. Mizutani, S., Yamada, T. & Yachida, S. Significance of the gut microbiome in multistep colorectal carcinogenesis. *Cancer Science* **111**, 766–773 (2020).

43. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat Commun* **6**, 6528 (2015).

44. Casasanta, M. A. *et al.* Fusobacterium nucleatum host-cell binding and invasion induces IL-8 and CXCL1 secretion that drives colorectal cancer cell migration. *Science Signaling* **13**, eaba9157 (2020).

45. Parhi, L. *et al.* Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. *Nat Commun* **11**, 3259 (2020).

46. Shao, W. *et al.* Conversion from epithelial to partial-EMT phenotype by Fusobacterium nucleatum infection promotes invasion of oral cancer cells. *Sci Rep* **11**, 14943 (2021).

47. Xu, C. *et al.* Fusobacterium nucleatum promotes colorectal cancer metastasis through miR-1322/CCL20 axis and M2 polarization. *Gut Microbes* **13**, 1980347 (2021).

48. Wu, Y. *et al.* Fn-Dps, a novel virulence factor of Fusobacterium nucleatum, disrupts erythrocytes and promotes metastasis in colorectal cancer. *PLOS Pathogens* **19**, e1011096 (2023).

49. Zhang, S. *et al.* Fusobacterium nucleatum promotes epithelial-mesenchymal transiton through regulation of the lncRNA MIR4435-2HG/miR-296-5p/Akt2/SNAI1 signaling pathway. *The FEBS Journal* **287**, 4032–4047 (2020).

50. Galeano Niño, J. L. *et al.* Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature* **611**, 810–817 (2022).

51. Routy, B. *et al.* Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).

52. Gao, Y. *et al.* Fusobacterium nucleatum enhances the efficacy of PD-L1 blockade in colorectal cancer. *Sig Transduct Target Ther* **6**, 1–10 (2021).

53. Gao, Y. *et al.* Fusobacterium nucleatum stimulates cell proliferation and promotes PD-L1 expression via IFIT1-related signal in colorectal cancer. *Neoplasia* **35**, 100850 (2023).

54. Jiang, S.-S. *et al.* Fusobacterium nucleatum-derived succinic acid induces tumor resistance to immunotherapy in colorectal cancer. *Cell Host Microbe* **31**, 781-797.e9 (2023).

55. Galeano Niño, J. L. *et al.* INVADEseq to identify cell-adherent or invasive bacteria and the associated host transcriptome at single-cell-level resolution. *Nat Protoc* **18**, 3355–3389 (2023).

56. Lo, C.-H. *et al.* Enrichment of Prevotella intermedia in human colorectal cancer and its additive effects with Fusobacterium nucleatum on the malignant transformation of colorectal adenomas. *Journal of Biomedical Science* **29**, 88 (2022).

57. Ternes, D. *et al.* The gut microbial metabolite formate exacerbates colorectal cancer progression. *Nat Metab* **4**, 458–475 (2022).

58. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**, 1350–1356 (2015).

59. Baran, B. *et al.* Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. *Gastroenterology Research* **11**, 264-273–273 (2018).

60. Kim, K., Castro, E. J. T., Shim, H., Advincula, J. V. G. & Kim, Y.-W. Differences Regarding the Molecular Features and Gut Microbiota Between Right and Left Colon Cancer. *Ann Coloproctol* **34**, 280–285 (2018).

61. Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S. & Frizelle, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Scientific Reports* **7**, (2017).

62. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).

63. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet* **48**, 1413–1417 (2016).

64. Lopera-Maya, E. A. *et al.* Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *Nat Genet* **54**, 143–151 (2022).

65. Zhernakova, D. V. *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* 1–9 (2024) doi:10.1038/s41586-023-06893-w.

66. Jasperson, K. W., Tuohy, T. M., Neklason, D. W. & Burt, R. W. Hereditary and Familial Colon Cancer. *Gastroenterology* **138**, 2044–2058 (2010).

67. Lynch, H. *et al.* Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet* **76**, 1–18 (2009).

68. Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of Lynch syndrome: 1895–2015. *Nature Reviews Cancer* **15**, 181–194 (2015).

69. Cerretelli, G., Ager, A., Arends, M. J. & Frayling, I. M. Molecular pathology of Lynch syndrome. *The Journal of Pathology* **250**, 518–531 (2020).

70. Bonadona, V. Cancer Risks Associated With Germline Mutations in MLH1, MSH2, and MSH6 Genes in Lynch Syndrome. *JAMA* **305**, 2304 (2011).

71. Ahadova, A. *et al.* Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. *International Journal of Cancer* **143**, 139–150 (2018).

72. Kastrinos, F. *et al.* Gene-Specific Variation in Colorectal Cancer Surveillance Strategies for Lynch Syndrome. *Gastroenterology* (2021) doi:10.1053/j.gastro.2021.04.010.

73. Capper, D. *et al.* BRAF V600E-specific immunohistochemistry for the exclusion of Lynch syndrome in MSI-H colorectal cancer. *International Journal of Cancer* **133**, 1624–1630 (2013).

74. Xiao, B. *et al.* Comparisons of screening strategies for identifying Lynch syndrome among patients with MLH1-deficient colorectal cancer. *Eur J Hum Genet* **28**, 1555–1562 (2020).

75. Wang, W. *et al.* A modified screening strategy for Lynch syndrome among MLH1-deficient CRCs: Analysis from consecutive Chinese patients in a single center. *Translational Oncology* **14**, 101049 (2021).

76. Battaglin, F., Naseem, M., Lenz, H.-J. & Salem, M. E. Microsatellite Instability in Colorectal Cancer: Overview of Its Clinical Significance and Novel Perspectives. *Clin Adv Hematol Oncol* **16**, 735–745 (2018).

77. Han, S. *et al.* The distinct clinical trajectory, metastatic sites, and immunobiology of microsatellite-instability-high cancers. *Front Genet* **13**, 933475 (2022).

78. Biller, L. H., Syngal, S. & Yurgelun, M. B. Recent advances in Lynch syndrome. *Familial Cancer* (2019) doi:10.1007/s10689-018-00117-1.

79. Walkowska, J. *et al.* Immunoprofiles of colorectal cancer from Lynch syndrome. *Oncoimmunology* **8**, e1515612 (2018).

80. Pastor, D. M. & Schlom, J. Immunology of Lynch Syndrome. *Curr Oncol Rep* **23**, 96 (2021).

81. Schwitalle, Y. *et al.* Immune Response Against Frameshift-Induced Neopeptides in HNPCC Patients and Healthy HNPCC Mutation Carriers. *Gastroenterology* **134**, 988–997 (2008).

82. Inamori, K. *et al.* Importance of lymph node immune responses in MSI-H/dMMR colorectal cancer. *JCI Insight* **6**, e137365 (2021).

83. Colle, R. *et al.* Parameters associated with outcomes in pretreated MSI/dMMR metastatic colorectal cancer (mCRC) treated with immune checkpoint inhibitors (ICI): Subgroup analysis of a prospective cohort. *JCO* **39**, 3580–3580 (2021).

84. Tominaga, T. *et al.* Pathological complete response to pembrolizumab in patients with metastatic ascending colon cancer with microsatellite instability. *Clin J Gastroenterol* (2021) doi:10.1007/s12328-021-01543-y.

85. Salman, P., Panay, S., Fernández, R., Mahave, M. & Soza-Ried, C. Evidence of response to pembrolizumab in a patient with Lynch syndrome-related metastatic colon cancer. *OncoTargets and Therapy* **Volume 11**, 7295–7300 (2018).

86. Hale, V. L. *et al.* Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. *Genome Med* **10**, 78 (2018).

87. Ono, T. *et al.* Fusobacterium nucleatum load in MSI colorectal cancer subtypes. *Int J Clin Oncol* **27**, 1580–1588 (2022).

88. Mima, K. *et al.* Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. *Gut* **65**, 1973–1980 (2016).

89. Tosti, E. *et al.* Loss of MMR and TGFBR2 increases the susceptibility to microbiota-dependent inflammation-associated colon cancer. *Cellular and Molecular Gastroenterology and Hepatology* S2352345X22000959 (2022) doi:10.1016/j.jcmgh.2022.05.010.

90. Alexander, M. *et al.* Human gut bacterial metabolism drives Th17 activation and colitis. *Cell Host & Microbe* **30**, 17-30.e9 (2022).

91. Chen, J., Pitmon, E. & Wang, K. Microbiome, inflammation and colorectal cancer. *Seminars in Immunology* **32**, 43–53 (2017).

92. Mori, G. *et al.* Gut Microbiota Analysis in Postoperative Lynch Syndrome Patients. *Front. Microbiol.* **10**, (2019).

93. Ferrarese, R. *et al.* Oral and Fecal Microbiota in Lynch Syndrome. *Journal of Clinical Medicine* **9**, 2735 (2020).

94. Rifkin, S. B. *et al.* Gut Microbiome Composition in Lynch Syndrome With and Without History of Colorectal Neoplasia and Non-Lynch Controls. *J Gastrointest Canc* (2023) doi:10.1007/s12029-023-00925-4.

95. Yan, Y. *et al.* Structure of the Mucosal and Stool Microbiome in Lynch Syndrome. *Cell Host & Microbe* **27**, 585-600.e4 (2020).

96. Tsugane, S. & Sawada, N. The JPHC Study: Design and Some Findings on the Typical Japanese Diet. *Japanese Journal of Clinical Oncology* **44**, 777–782 (2014).

97. Soga, T. *et al.* Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry. *J. Proteome Res.* **2**, 488–494 (2003).

98. Hirayama, A. *et al.* Quantitative Metabolome Profiling of Colon and Stomach Cancer Microenvironment by Capillary Electrophoresis Time-of-Flight Mass Spectrometry. *Cancer Res* **69**, 4918–4925 (2009).

99. Erawijantari, P. P. *et al.* Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut* **69**, 1404–1415 (2020).

100. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).

101. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

102. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research* **42**, D643–D648 (2014).

103. Mori, H., Maruyama, T., Yano, M., Yamada, T. & Kurokawa, K. VITCOMIC2: visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing. *BMC Systems Biology* **12**, 30 (2018).

104. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

105. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

106. Besemer, J. & Borodovsky, M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Research* **27**, 3911–3920 (1999).

107. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587–D592 (2023).

108. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).

109. Umaña, A. *et al.* Utilizing Whole Fusobacterium Genomes To Identify, Correct, and Characterize Potential Virulence Protein Families. *Journal of Bacteriology* **201**, (2019).

110. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).

111. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).

112. Dougherty, M. W. *et al.* The microbial genotoxin colibactin exacerbates mismatch repair mutations in colorectal tumors. *Neoplasia* **43**, 100918 (2023).

113. Nougayrède, J.-P. *et al. Escherichia coli* Induces DNA Double-Strand Breaks in Eukaryotic Cells. *Science* **313**, 848–851 (2006).

114. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).

115. Velez-Cortes, F. & Wang, H. Characterization and Spatial Mapping of the Human Gut Metasecretome. *mSystems* **7**, e00717-22 (2022).

116. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *European Journal of Cell Biology* **97**, 422–441 (2018).

117. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* **40**, 1023–1025 (2022).

118. Glover, J. S., Ticer, T. D. & Engevik, M. A. Characterizing the mucin-degrading capacity of the human gut microbiota. *Sci Rep* **12**, 8456 (2022).

119. Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**, 1339-1353.e21 (2016).

120. Berkhout, M. D., Plugge, C. M. & Belzer, C. How microbial glycosyl hydrolase activity in the gut mucosa initiates microbial cross-feeding. *Glycobiology* **32**, 182–200 (2021).

121. Labarthe, S. *et al.* Four functional profiles for fibre and mucin metabolism in the human gut microbiome. *Microbiome* **11**, 231 (2023).

122. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* **50**, D571–D577 (2022).

123. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**, W95–W101 (2018).

124. Xia, Y. & Sun, J. Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases* **4**, 138–148 (2017).

125. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).

126. Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nat Microbiol* **3**, 8–16 (2018).

127. Holmes, I., Harris, K. & Quince, C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLOS ONE* **7**, e30126 (2012).

128. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 1–27 (2021).

129. Ghosh, T. S., Das, M., Jeffery, I. B. & O'Toole, P. W. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. (2020) doi:10.7554/eLife.50240.

130. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLOS Computational Biology* **17**, e1009442 (2021).

131. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).

132. Olovo, C. V., Huang, X., Zheng, X. & Xu, M. Faecal microbial biomarkers in early diagnosis of colorectal cancer. *J Cell Mol Med* **25**, 10783–10797 (2021).

133. Rynazal, R. *et al.* Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biology* **24**, 21 (2023).

134. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).

135. Enteropathway: the metabolic pathway database for the human gut microbiota | bioRxiv. https://www.biorxiv.org/content/10.1101/2023.06.28.546710v1.

136. Naddaf, R. *et al.* Gut microbial signatures are associated with Lynch syndrome (LS) and cancer history in Druze communities in Israel. *Sci Rep* **13**, 20677 (2023).

137. Ning, L. *et al.* Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. *Nat Commun* **14**, 7135 (2023).

138. Bohaumilitzky, L. *et al.* The Different Immune Profiles of Normal Colonic Mucosa in Cancer-Free Lynch Syndrome Carriers and Lynch Syndrome Colorectal Cancer Patients. *Gastroenterology* **162**, 907-919.e10 (2022).

139. Chang, K. *et al.* Immune Profiling of Premalignant Lesions in Patients With Lynch Syndrome. *JAMA Oncol* **4**, 1085–1092 (2018).

140. Ten Broeke, S. W. *et al.* Cancer Risks for PMS2-Associated Lynch Syndrome. *J Clin Oncol* **36**, 2961–2968 (2018).

141. Liu, Y. L. *et al.* Early age of onset and broad cancer spectrum persist in MSH6- and PMS2-associated Lynch syndrome. *Genetics in Medicine* **24**, 1187–1195 (2022).

142. Abed, J. *et al.* Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host & Microbe* **20**, 215–225 (2016).

143. Gur, C. *et al.* Binding of the Fap2 Protein of Fusobacterium nucleatum to Human Inhibitory Receptor TIGIT Protects Tumors from Immune Cell Attack. *Immunity* **42**, 344–355 (2015).

144. Galaski, J. *et al.* Fusobacterium nucleatum CbpF Mediates Inhibition of T Cell Function Through CEACAM1 Activation. *Frontiers in Cellular and Infection Microbiology* **11**, (2021).

145. Engevik, M. A. *et al.* Fusobacterium nucleatum Adheres to Clostridioides difficile via the RadD Adhesin to Enhance Biofilm Formation in Intestinal Mucus. *Gastroenterology* **160**, 1301-1314.e8 (2021).

146. Kaplan, C. W. *et al.* Fusobacterium nucleatum Outer Membrane Proteins Fap2 and RadD Induce Cell Death in Human Lymphocytes. *Infect Immun* **78**, 4773–4778 (2010).

147. Nosho, K. *et al.* Association of Fusobacterium nucleatum with immunity and molecular alterations in colorectal cancer. *World Journal of Gastroenterology* **22**, 557–566 (2016).

148. Guo, P. *et al.* FadA promotes DNA damage and progression of Fusobacterium nucleatum-induced colorectal cancer through up-regulation of chk2. *Journal of Experimental & Clinical Cancer Research* **39**, 202 (2020).

149. Hamada, T. *et al.* Fusobacterium nucleatum in Colorectal Cancer Relates to Immune Response Differentially by Tumor Microsatellite Instability Status. *Cancer Immunol Res* **6**, 1327–1336 (2018).

150. Lee, J. A. *et al.* Differential immune microenvironmental features of microsatellite-unstable colorectal cancers according to Fusobacterium nucleatum status. *Cancer Immunol Immunother* **70**, 47–59 (2021).

151. Kuo, C.-C., Wu, J.-Y. & Wu, K. K. Cancer-derived extracellular succinate: a driver of cancer metastasis. *Journal of Biomedical Science* **29**, 93 (2022).

152. Li, H. *et al.* Asymmetric dimethylarginine attenuates serum starvation-induced apoptosis via suppression of the Fas (APO-1/CD95)/JNK (SAPK) pathway. *Cell Death Dis* **4**, e830–e830 (2013).

153. Gobert, A. P. *et al.* Protective Role of Spermidine in Colitis and Colon Carcinogenesis. *Gastroenterology* **162**, 813-827.e8 (2022).

154. Ding, N. *et al.* Fusobacterium nucleatum Infection Induces Malignant Proliferation of Esophageal Squamous Cell Carcinoma Cell by Putrescine Production. *Microbiology Spectrum* **11**, e02759-22 (2023).

155. Liu, Y. *et al.* Secondary Bile Acids and Tumorigenesis in Colorectal Cancer. *Frontiers in Oncology* **12**, (2022).

156. Poetsch, A. R. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Computational and Structural Biotechnology Journal* **18**, 207–219 (2020).

157. Xia, L. *et al.* The cancer metabolic reprogramming and immune response. *Molecular Cancer* **20**, 28 (2021).