

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**DETERMINATION OF ENTEROTYPE SPECIFIC MICROBIAL
BIOMARKERS FROM METAGENOME DATA FOR EARLY DIAGNOSIS
AND SCREENING SYSTEM IN COLON CANCER**

ÜNZİLE GÜVEN GÜLHAN

A THESIS OF DOCTORATE
BIOENGINEERING DEPARTMENT
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAM

ADVISOR: DR. SALİHA DURMUŞ

JULY 2024

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**DETERMINATION OF ENTEROTYPE SPECIFIC MICROBIAL
BIOMARKERS FROM METAGENOME DATA FOR EARLY DIAGNOSIS
AND SCREENING SYSTEM IN COLON CANCER**

ÜNZİLE GÜVEN GÜLHAN

**A THESIS OF DOCTORATE
BIOENGINEERING DEPARTMENT
BIOINFORMATICS AND SYSTEMS BIOLOGY
PROGRAM**

ADVISOR: DR. SALİHA DURMUŞ

JULY 2024

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

KOLON KANSERİNDE ERKEN TANI VE TARAMA SİSTEMİ İÇİN
METAGENOM VERİLERİNDEN ENTEROTİP SPESİFİK MİKROBİYAL
BİYOBELİRTEÇ TAYİNİ

ÜNZİLE GÜVEN GÜLHAN

DOKTORA TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ
PROGRAMI

DANIŞMAN: DR. SALİHA DURMUŞ

TEMMUZ 2024

DOCTORATE JURY APPROVAL FORM

A thesis submitted by Ünzile Güven Gülhan, defended on 24/07/2024 before the jury formed with the 04/07/2024 date and 2024/34 numbered decision of the GTU Graduate Administration Board, has been accepted as a DOCTOR of PHILOSOPHY thesis in the Department of Bioengineering, Bioinformatics and Systems Biology Program.

JURY

MEMBER

(THESIS ADVISOR) : Dr. Saliha DURMUŞ

MEMBER : Prof. Dr. Tunahan ÇAKIR

MEMBER : Assoc. Prof. Dr. Emrah NİKEREL

MEMBER : Prof. Dr. Fatih Erdoğan SEVİLGİN

MEMBER : Assist. Prof. Dr. Pınar PİR

APPROVAL

Gebze Technical University Graduate Administration Board
...../...../..... date and/..... numbered decision.

SIGNATURE/SEAL

*To my Family and Friends, whose
unwavering support and encouragement have been
invaluable throughout this journey.*

ABSTRACT

Colorectal cancer (CRC) ranks second among cancer deaths in the world. It is known that colon cancer cases can be prevented to a great extent if diagnosed at an early stage. Existing non-invasive colon cancer screening tests are not sufficient to detect early phase tumor formation. As a result of the accelerated experimental studies in recent years, it has been observed that human intestinal microbiota is associated with colon cancer, and it has been shown that changing microbiome profiles in human microbiota can be used as biomarkers. In this context, colon cancer-human microbiota interaction and related microbial composition changes were investigated by metagenomic analysis. We used publicly available human fecal samples obtained from whole genome shotgun metagenome sequencing (N=1579) and 16S rRNA amplicon sequencing (N=1949) to investigate microbiome-CRC associations. Analyses of these data were performed using HUMAnN2 and QIIME2 pipelines, revealing taxonomic and functional compositions of the gut microbiota in healthy controls (CTR), patients with CRC and those with adenoma (ADN). Gut microbiota of healthy people were clustered into microbial subtypes and enterotypes were determined. Enterotypes have been shown to be an important factor for population stratification based on the gut microbiota composition, leading to a better understanding of human health and disease states. Classifications based on compositional patterns will have implications for personalized microbiota-based solutions. There has been limited enterotype based studies on colorectal adenoma and cancer. We performed an enterotype-based meta-analysis of the publicly available metagenomic studies and gut microbiota of healthy people were clustered into three enterotypes (Ruminococcus-, Bacteroides- and Prevotella dominated enterotypes). Reference-based enterotype assignments were performed for CRC and ADN samples, using the supervised machine learning algorithm, K-Nearest Neighbors. Differential abundance analyses and Random Forest classification were conducted on each enterotype between healthy controls and CRC-ADN groups, revealing novel enterotype-specific microbial markers for non-invasive CRC screening strategies. We also identified microbial genes and pathways that undergo significant changes in CRC and ADN for each enterotype. Furthermore, the potential of bioactive molecules to target significantly changed microbial genes in the metagenome of cancer patients were also evaluated to identify therapeutic opportunities and suggest drug repositioning strategies. This study is promising in elucidating the mechanisms of colon cancer and adenoma-to-carcinoma sequence under different gut microbiome profiles.

Keywords: Gut microbiota, Colorectal cancer, Adenoma, Biomarker, Metagenomics, Enterotypes

ÖZET

Kolorektal kanser (KRK), dünyada kanser ölümleri arasında ikinci sırada yer almaktadır. Kolon kanseri vakalarının erken evrede teşhis edilmesi durumunda büyük ölçüde önlenilebileceği bilinmektedir. Mevcut non-invaziv kolon kanseri tarama testleri, erken evre tümör oluşumunu tespit etmek için yeterli değildir. Son yıllarda hızlanan deneysel çalışmalar sonucunda, insan bağırsak mikrobiyotasının kolon kanseri ile ilişkili olduğu gözlemlenmiş ve insan mikrobiyotasındaki değişen mikrobiyom profillerinin biyomarker olarak kullanılabileceği gösterilmiştir. Bu bağlamda, metagenomik analiz ile kolon kanseri-insan mikrobiyota etkileşimi ve ilgili mikrobiyal kompozisyon değişiklikleri araştırılmıştır. Mikrobiyom-KRK ilişkilerini incelemek için, tüm genom shotgun metagenom dizileme (N=1579) ve 16S rRNA amplikon dizileme (N=1949) ile elde edilen halka açık insan dışkı örnekleri kullanılmıştır. Bu verilerin analizi HUMAnN2 ve QIIME2 iş akışları kullanılarak gerçekleştirilmiş ve sağlıklı kontroller (CTR), KRK hastaları (CRC) ve adenoma (ADN) hastalarının bağırsak mikrobiyota taksonomik ve fonksiyonel kompozisyonları ortaya çıkarılmıştır. Sağlıklı bireylerin bağırsak mikrobiyotası mikrobiyal alt türlere ayrılmış ve enterotipler belirlenmiştir. Enterotiplerin, bağırsak mikrobiyota kompozisyonuna dayalı olarak popülasyon sınıflandırmasında önemli bir faktör olduğu ve insan sağlığı ve hastalık durumlarının daha iyi anlaşılmasını sağladığı gösterilmiştir. Kompozisyonel modellere dayalı sınıflandırmalar, kişiselleştirilmiş mikrobiyota tabanlı çözümler için etkili olacaktır. Kolorektal adenom ve kanser üzerine sınırlı enterotip bazlı çalışmalar mevcuttur. Bu tez kapsamında, halka açık metagenomik çalışmaların enterotip bazlı meta-analizi gerçekleştirilmiş ve sağlıklı bireylerin bağırsak mikrobiyotası üç enterotipe (Ruminococcus-, Bacteroides- ve Prevotella baskın enterotipler) ayrılmıştır. Referans tabanlı enterotip atamaları, K-En Yakın Komşu adlı denetimli makine öğrenme algoritması kullanılarak CRC ve ADN örnekleri için gerçekleştirilmiştir. Sağlıklı kontroller ve CRC-ADN grupları arasında her enterotip için farklı miktar analizleri ve Random Forest sınıflandırması yapılmış, non-invaziv KRK tarama stratejileri için yeni enterotip-spesifik mikrobiyal belirteçler ortaya konmuştur. Ayrıca, her enterotip için KRK ve ADN'de önemli değişiklikler gösteren mikrobiyal genler ve yollar belirlenmiştir. Ek olarak, kanser hastalarının metagenomunda önemli değişiklikler gösteren mikrobiyal genleri hedef almak için biyolojik aktif moleküllerin potansiyeli değerlendirilmiş, terapötik fırsatlar ve ilaç yeniden konumlandırma stratejileri önerilmiştir. Bu çalışma, farklı bağırsak mikrobiyom profilleri altında kolon kanseri ve adenomdan karsinoma aşamasına geçiş mekanizmalarını aydınlatmada umut vaat etmektedir.

Anahtar Kelimeler: Bağırsak mikrobiyotası, Kolorektal kanser, Adenom, Biyobelirteç, Metagenomik, Enterotipler

ACKNOWLEDGEMENTS

To my family, thank you for your unwavering love and belief in me. Your sacrifices and support were my foundation throughout this journey. Special thanks to my mother-in-law for her invaluable support in raising my son, which allowed me to focus on my studies.

I am profoundly thankful to my advisor, Dr. Saliha Durmuş and the members of my dissertation committee, Prof. Dr. Tunahan Çakır, Prof. Dr. Fatih Erdoğan Sevilgen, Assoc. Prof. Dr. Emrah Nikerel, for their support and invaluable guidances throughout my Ph.D. journey.

I would also like to acknowledge the financial support from TÜSEB and PhiTech Bioinformatics. Their contributions provided the necessary resources and funding to pursue and complete this research.

A special thanks to my colleagues and friends in the Bioinformatics Department of PhiTech Bioinformatics, whose fellowship and collaboration were truly encouraging.

Lastly, to my son Ahmet Akif and to my husband Erman, your presence in my life is a constant source of happiness and gratitude.

This thesis is dedicated to my family, whose love and understanding made this achievement possible.

Thank you all for being part of this journey.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xi
LIST OF FIGURES	xii
LIST OF TABLES	xiv
1. INTRODUCTION	1
2. BACKGROUND ASPECTS	3
2.1. Advances in Human Microbiome Research and its Implications for Cancer Studies	3
2.2. Biomarkers with Potential Use in CRC and Precursor Adenoma Screening	4
2.3. Enterotypes and Their Role in Colorectal Cancer Biomarkers	6
2.4. Biomarkers Use of Intestinal Microbiota and Microbial Gene-Targeted Drug Approach to CRC Treatment	7
3. MATERIALS AND METHODS	9
3.1. Public Data Collection	9
3.1.1. Public Data Collection of Whole Genome Shotgun Metagenomes	9
3.1.2. Public Data Collection of 16S rRNA Amplicon Sequencing	9
3.2. Data Preprocessing and Taxonomic Profiling	9
3.2.1. Data Preprocessing and Taxonomic Profiling for WGS Dataset	9
3.2.2. Data Preprocessing and Taxonomic Profiling for 16S Dataset	10
3.3. Enterotype Clustering and Classification	11
3.4. Confounder Analyses Based on Metadata	12
3.5. Identification of Enterotype Specific Microbial Species Using Differential Abundance Analysis (DAA)	12
3.6. Model Construction and Identification of ET-Specific Microbial Markers	13
3.7. Microbial Co-occurrence Analysis	14
3.8. Functional Gene Family and Pathway Analysis of Enterotypes	14
3.9. Correlation Analysis of Microbial Markers for CRC and ADN	15
3.10. Microbial Gene/Protein Targeted Drug Screening	15
4. RESULTS	16
4.1. Datasets Characteristics in Meta-analysis	16
4.2. Enterotype Identification	17
4.3. Identification of Potential Confounders	20
4.4. ET-specific Differential Microbial Species	23
4.5. Model Construction and Identification of ET-Specific Microbial Markers	33
4.6. Enterotype-Specific Co-occurrence Networks	48

4.7. Functional Gene Families and Pathways Associated with ET-Specific Microbial Markers	52
4.8. Correlation Analysis of ET-Specific Microbial Markers Associated with BMI and Age	60
4.9. Microbial Gene/Protein Targeted Drug Screening	62
5. DISCUSSION	63
REFERENCES	71
BIOGRAPHY	84
PUBLICATIONS AND PRESENTATIONS FROM THE THESIS	85
APPENDICES	86

LIST OF SYMBOLS AND ABBREVIATIONS

16S	: 16S rDNA Amplicon Sequencing
ADN	: Adenoma
ANCOM-BC	: Analysis of Compositions of Microbiomes with Bias Correction
ASV	: Amplicon Sequence Variant
AUC	: Area Under the Curve
BMI	: Body Mass Index
CH	: Calinski-Harabasz
CRC	: Colorectal Cancer
CTR	: Control
DAA	: Differential Abundance Analysis
eggNOG	: evolutionary gene genealogy Non-supervised Orthologous Groups
ET	: Enterotype
FIT	: Fecal Immunochemical Test
FMT	: Fecal Microbiome Transfer
IBD	: Inflammatory Bowel Diseases
iHMP	: Integrative Human Microbiome Project
JSD	: Jensen–Shannon Divergence
KEGG	: Kyoto Encyclopedia of Genes and Genomes
KNN	: K-Nearest Neighbors
LDA	: Linear Discriminant Analysis
LEfSe	: Linear Discriminant Analysis Effect Size
MetaPhlAn	: Metagenomic Phylogenetic Analysis
NIH	: National Institutes of Health
PAM	: Partitioning Around Medoids
PCoA	: Principal-Coordinate Analysis
QIIME	: Quantitative Insights Into Microbial Ecology
RF	: Random Forest
ROC	: Receiver Operating Characteristic
SH	: Silhouette Coefficient
SIAMCAT	: Statistical Inference of Associations between Microbial Communities And host phenoTypes
SSU	: Small SubUnit
T2D	: Type 2 Diabetes
UniRef	: The UniProt Reference Clusters
WGS	: Whole Genome shotgun metagenome Sequencing

LIST OF FIGURES

	<u>Page</u>
Figure 4.1: PCoA plot generated using the PAM clustering method based on Jensen- Shannon distance of microbial abundances for A) WGS dataset and B) 16S dataset. 1) Eubacterium-Ruminococcus (Blautia) dominated enterotype cluster (ET1), 2) Bacteroides dominated enterotype cluster (ET2), 3) Prevotella dominated enterotype cluster (ET3)	19
Figure 4.2: Variance explained by disease status (Control, CRC, ADN) in WGS dataset is plotted against variance explained by metadata variables (Age, BMI, Sex, Study) for microbial species in A) Enterotype 1, B) Enterotype 2 and C) Enterotype 3	21
Figure 4.3: Variance explained by disease status (Control, CRC, ADN) in 16S dataset is plotted against variance explained by metadata variables (Age, BMI, Sex, Study) for microbial species in A) Enterotype 1, B) Enterotype 2 and C) Enterotype 3	22
Figure 4.4: Comparison of differentially abundant microbial species for A) CRC-CTR, B) ADN-CTR and c) ADN-CRC comparison in each 3 enterotypes detected using WGS dataset.	24
Figure 4.5: Comparison of differentially abundant microbial species for A) CRC-CTR, B) ADN-CTR and c) ADN-CRC comparison in each 3 enterotypes detected using 16S dataset.	25
Figure 4.6: The mean contribution of top 10 marker species to the classification between CRC and CTR in each enterotype.	34
Figure 4.7: ROC curves of the Random Forest models for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.	35
Figure 4.8: Heatmap visualizations of the Random Forest Classifier results for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.	37
Figure 4.9: Venn diagram comparing Enterotype (ET) specific and non-ET specific CRC markers detected using WGS dataset.	37
Figure 4.10: The mean contribution of top 10 marker species to the classification between CRC and CTR in each enterotype, using 16S dataset	38
Figure 4.11: ROC curves of the Random Forest models for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.	39
Figure 4.12: Heatmap visualizations of the Random Forest Classifier results for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.	40
Figure 4.13: Venn diagram comparing ET specific ADN markers in the adenoma-carcinoma sequence	41
Figure 4.14: ROC curves of the Random Forest models for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.	42

Figure 4.15: The mean contribution of top 10 marker species to the classification between CRC and ADN in each enterotype, using WGS dataset	42
Figure 4.16: Heatmap visualizations of the Random Forest Classifier results for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.	44
Figure 4.17: The mean contribution of top 10 marker species to the classification between CRC and ADN in each enterotype, using 16S dataset	45
Figure 4.18: ROC curves of the Random Forest models for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.	45
Figure 4.19: Heatmap visualizations of the Random Forest Classifier results for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.	47
Figure 4.20: Microbial co-occurrence network plots for Enterotypes detected using WGS dataset.	50
Figure 4.21: Comparison of the top 20 most effective (LDA>2) pathways in ETs associated with A) CRC and B) ADN	60

LIST OF TABLES

	<u>Page</u>
Table 4.1: Characteristics of the fecal metagenomic datasets included in this study	16
Table 4.2: Characteristics of the fecal 16S datasets included in this study	17
Table 4.3: Characteristics of the fecal WGS datasets included in this study	17
Table 4.4: Characteristics of the fecal 16S datasets included in this study	18
Table 4.5: Number of WGS samples assigned to each enterotype by clustering/classification analysis	19
Table 4.6: Number of 16S samples assigned to each enterotype by clustering/classification analysis	20
Table 4.7: ET-specific differential species related to CRC, detected from WGS dataset	25
Table 4.8: ET-specific differential species related to ADN when compared to CTR detected from WGS dataset	27
Table 4.9: ET-specific differential species related to ADN when compared to CRC detected from WGS dataset	28
Table 4.10: ET-specific differential species related to CRC when compared to CTR detected from 16S dataset	30
Table 4.11: ET-specific differential species related to ADN when compared to CTR detected from 16S dataset	31
Table 4.12: ET-specific differential species related to ADN when compared to CRC detected from 16S dataset	32
Table 4.13: Co-abundant microbial species detected in CRC, CTR and ADN samples classified into three ETs.	51
Table 4.14: ET-specific CRC-associated genes and their species	53
Table 4.15: ET-specific ADN-associated genes and their species	55
Table 4.16: The top 10 differentially expressed ET-specific pathways	57
Table 4.17: Results with high correlation between age, BMI and microbial markers	61

1. INTRODUCTION

Colorectal cancer (CRC) is one of the most common types of cancer worldwide, accounting for approximately 10% of all cancer cases. It is a major cause of morbidity and mortality, with an estimated 1.9 million new cases and 935,000 deaths in 2020. Adenoma (ADN), a precursor lesion of CRC, is also highly prevalent and is estimated to be present in approximately 20-30% of adults over 50 years of age (Brenner et al., 2014; Dekker et al., 2019; Hewitson et al., 2007). Dysbiosis, or an imbalance in the gut microbiota composition, has been associated with CRC and ADN, with studies showing that healthy individuals have distinct microbial communities compared to those with CRC or ADN (Ponziani et al., 2019; Yu et al., 2017). As a result of these studies, the use of intestinal microbiota as a biomarker in the clinic has increased its importance in recent years and it is stated that the potential for colon cancer diagnosis and screening will be high. However, the functional mechanisms underlying human and microbiome interaction for many diseases, including colon cancer, have not been fully elucidated, and the implementation of holistic functional analyzes in this area will contribute greatly.

Enterotypes are distinct clusters of gut microbiota including clusters dominated by *Bacteroides*, *Prevotella* and *Ruminococcus* (*Eubacteriales* order) species (Arumugam et al., 2011) which have demonstrated associations with long-term dietary patterns (Wu et al., 2011a). Utilizing enterotypes for stratifying individuals into groups based on their gut microbiota composition has become a significant aspect of population stratification and the development of personalized microbiota-based solutions. There are studies in which compositional differences in certain enterotypes are associated with different diseases (depression, obesity, cancer, etc.) (Bushman et al., 2013; Valles-Colomer et al., 2019; Vieira-Silva et al., 2019). However, there has been limited research on the enterotype-specific microbial features in CRC and ADN (Claesson et al., 2012; Costea et al., 2018; Wu et al., 2011a). Microbial compositional patterns can aid in classification of populations that will serve the purposes of precision medicine, leading to more precise prediction, diagnosis and treatments. Such classifications are commonly used in molecular subtyping of complex diseases like cancers, where clinically relevant patterns (like gene expression patterns) are used to determine

appropriate therapeutic intervention for patients (Sørli et al., 2001, 2003). There is increasing interest in the usefulness of such classifications in colorectal cancer cases (Khaliq et al., 2022; D. Lin et al., 2021). The development of subtype-specific markers may contribute to more effective management of this challenging disease. However, molecular stratification may not be applicable in all situations since CRC is a highly heterogeneous disease (Costea et al., 2018; Yamauchi et al., 2012). The gut microbial ecosystem has been shown to be an important factor in disease progression from adenoma to cancer (Kim & Lim, 2022) and classification of gut communities by enterotypes has the potential to be clinically useful (Costea et al., 2018). With enterotype-based analysis, more specific markers can be discovered and it is possible to measure CRC risk factors in people with certain enterotypes (Christensen et al., 2018; Costea et al., 2018). However, little is known about the effect of changes in different enterotypes on colorectal cancer and precursor adenomas. Also, there is no study suggesting enterotype-specific markers at species resolution for early diagnosis markers of CRC.

In this thesis study, published metagenome data were collected from patients with CRC and have adenomatous polyps, along with healthy people's metagenomes to be used as the control. Healthy individuals' enterotype clusters with the patients' microorganism communities that fall in the same cluster were compared. Through this approach, the differential communities in cancerous and adenomatous stages were observed, suggesting them as enterotype-specific disease markers. Additionally, various mechanisms occurring within these enterotype clusters were discussed in alignment with dietary patterns. This study provided significant insights for further investigations leading to the development of non-invasive CRC diagnostic tools and therapeutic interventions.

2. BACKGROUND ASPECTS

2.1. Advances in Human Microbiome Research and its Implications for Cancer Studies

With the first phase of the Human Microbiome Project, which started in 2007 and ended in 2012, more than 2000 species were found as a result of metagenome analyses performed in 300 healthy people (Turnbaugh et al., 2007). With the second phase of the project (iHMP) between 2013-2016, functional analyses were carried out to determine how this microbiota profile is affected in different health conditions and the existence of the relationship of microbiome with a specific phenotype. Preliminary studies were conducted by selecting 3 different conditions with iHMP, and the relationship between preterm birth and maternal vaginal microbiome, Inflammatory Bowel Diseases (IBD) and Type 2 Diabetes (T2D) and human microbiota were elucidated (NIH Human Microbiome Portfolio Analysis Team, 2019). After the discovery of the effects of human microbiota on human health, analysis of its effects on cancer has come to the fore.

Human microbiota studies have focused mostly on the intestinal microbiota, which cover the majority of the microbiome, and many analyzes have been performed on easy-to-access fecal samples. The findings of these studies have led to the clinical importance of the use of microbiota associated with the disease as a marker (Mohajeri et al., 2018). A marker plays an indicator role in the diagnosis or determination of a disease. Many studies on colorectal cancer (CRC) have emphasized the importance of microbial markers and the potential for increased clinical use (Wong et al., 2019).

Numerous metagenomes and amplicon sequencing studies have been performed using the next generation sequencing technique on CRC and human microbiota correlation. Although there are many data sets related to this subject in public databases, studies have been conducted with limited populations and working groups. However, it is known that the microbiota profiles associated with CRC in the populations may be different, but we do not know enough about this variation yet. Although there are meta-analysis studies conducted with several population groups (Thomas et al., 2019;

Wirbel et al., 2019a), there are different microbial markers in different meta-analysis results. One of the reasons why there is no general stabilization in this regard is the limited number of studies conducted with very large groups of different populations. This thesis aims to conduct a meta-analysis in a large collection of datasets (more than 1000 samples from different populations), thus potentially will lead the way to effective biomarker explorations.

Since colon cancer is largely heterogeneous, markers customized for different types of cancer will have higher efficacy in diagnosis and treatment. For this reason, studies on cancer patients in the early stage are important for colon cancer screenings. CRC scans on risky patients have shown that CRC cases and mortality are reduced (Lauby-Secretan et al., 2018). Biomarkers identified as a result of analysis with this disease group will reduce the risk of death by controlling the disease at an early stage for CRC. For this purpose, there are studies in literature with different patients and cancer groups. Colorectal adenoma is known to be the precursor of approximately 90-95% of cases of colon cancer (Strum, 2016). Many cases of CRC are caused by the conversion of benign polyps to malignant forms, which is why CRC screening for adenoma patients with high cancer risk is crucial. Current CRC screening studies do not provide sufficient specificity and efficacy (Wong et al., 2019). Metagenome studies have accelerated and increased its importance to find more effective biomarkers with the discovery that human microbiota is critical for the onset, course and development of the disease (Saus et al., 2019). Increased metagenomic datasets for colon cancer provide a rich resource for the development of fecal microbial markers for screening and diagnostic purposes.

2.2. Biomarkers with Potential Use in CRC and Precursor Adenoma Screening

Early diagnosis of colon cancer is critical in terms of treatability and achieving positive clinical outcomes. For example, the 5-year survival rate in CRC cases that are not in the metastatic stage is 90%, it decreases to 14% for CRC in the metastatic stage (Siegel et al., 2019).

The sensitivity of fecal immunochemical test (FIT) tests to detect CRC, which is one of the non-invasive techniques used for CRC screening, remains at 79%, whereas this sensitivity rate decreases to 25-27% in adenoma cases (Lee et al., 2014). Stool DNA

testing, another screening test, could potentially detect more cancers than FIT (sensitivity 92.3%), but the sensitivity rate to detect advanced adenoma remains at 42.4% (Imperiale et al., 2014). Therefore, there is a need to develop a non-invasive screening test with a high accuracy and high sensitivity for both CRC and adenoma.

Different marker microorganisms were identified as a result of metagenome analysis aimed at finding microbial biomarkers that differentiate between colon cancer patients and control samples. In one study, there were 15 bacterial markers (Feng et al., 2015) in 41 Australian patients, whereas in another study, 32 bacterial markers were reported in 101 American patients (Baxter et al., 2016). In a meta-analysis with Italian, German and Japanese patient samples, 16 bacterial markers were found (Thomas et al., 2019). With the study with only 54 Japanese patients 54 bacterial markers were reported (Yachida et al., 2019). Metagenome analysis of 74 Danish and 16 Chinese patients resulted in a list of 20 microbial genes associated with colon cancer. As a result of functional analysis, this list has been reduced to two as *rpoB* gene of *P. micra* bacterium and Butyryl-CoA Dehydrogenase enzyme of *Fusobacterium nucleatum* bacterium. In another study of 30 American patients, 16S rRNA amplicon sequence analysis revealed 6 bacterial markers with 80% test accuracy and sensitivity rate increased to 92% when age, race and BMI factors were considered (Zackular et al., 2014). As a result of many studies, different markers have been reported for colon cancer screening and it has been found that the most common bacteria belong to the genus *Fusobacterium*. In a meta-analysis published in 2019, 7 different study groups were tested and 5 of them were reported as potential markers of *F. nucleatum* bacteria (Thomas et al., 2019). Analysis of screening tests in patients with adenoma indicated that the potential of *F. nucleatum* bacteria to be a marker, but less distinctive (Wong et al., 2019). In addition, while 5 bacterial markers were reported to have 90% diagnostic power (Zackular et al., 2014), it was reported that this rate was 58% in another study group (Thomas et al., 2019).

As can be seen, there are no common microbial compositions to be used in the fecal diagnosis and screening tests of colon cancer. The reasons for this may be the inability to collect samples or data with sufficient ethnic diversity, the use of different analysis methods and software tools, and the inclusion of clinical parameters.

2.3. Enterotypes and Their Role in Colorectal Cancer Biomarkers

Potential Enterotypes, distinct clusters of gut microbiota composition, have emerged as a significant factor in understanding and diagnosing colorectal cancer (CRC). These enterotypes are primarily dominated by specific bacterial genera such as *Bacteroides*, *Prevotella*, and *Ruminococcus* (Arumugam et al., 2011). The classification of gut microbiota into enterotypes allows researchers to stratify populations based on their microbial composition, which is crucial for personalized microbiota-based solutions (Costea et al., 2018).

Recent studies have demonstrated that enterotypes are associated with long-term dietary patterns, which influence the gut microbiota composition and, consequently, the risk of developing CRC (Wu et al., 2021). By analyzing the microbiota profiles of healthy individuals and comparing them with those of CRC and adenoma patients, researchers can identify enterotype-specific microbial markers that are indicative of disease presence and progression (Ponziani et al., 2019; Yu et al., 2017).

Enterotype-based analysis of gut microbiota has revealed distinct microbial communities in different enterotypes, which can serve as biomarkers for CRC. For instance, certain microbial species and functional pathways may be enriched or depleted in specific enterotypes of CRC patients compared to healthy controls (Lin et al., 2021). These findings suggest that enterotype-specific biomarkers could improve the accuracy and efficacy of non-invasive CRC screening methods (Khaliq et al., 2022).

Furthermore, the identification of enterotype-specific changes in the gut microbiota can provide insights into the underlying mechanisms of CRC development and progression. This knowledge can lead to the development of targeted therapies and personalized treatment strategies based on an individual's enterotype, enhancing the overall effectiveness of CRC management (Christensen et al., 2018).

In summary, enterotypes play a crucial role in the identification of colorectal cancer biomarkers. By leveraging the distinct microbial compositions of different enterotypes, researchers can develop more precise and personalized approaches for CRC diagnosis, screening, and treatment, ultimately improving patient outcomes (Vieira-Silva et al., 2019).

2.4. Biomarkers Use of Intestinal Microbiota and Microbial Gene-Targeted Drug Approach to CRC Treatment

Potential clinical applications of human microbiota in the treatment of CRC include determining the changing microbiome composition structure in cancer patients and modifying the microbiota structure to produce these microorganisms. Clinical trials initiated using this approach are thought to be promising (Vivarelli et al., 2019). This therapy approach is mainly aimed at increasing the effects of chemotherapy and immunotherapy drugs and improving the response to treatment.

Another potential clinical application would be the development of novel drug molecules that inhibit mechanisms associated with intestinal microbiota causing colon cancer. In order to elucidate these mechanisms, it is important to uncover the clusters and network structures of microorganisms that are detected in high amounts in cancer. If the resulting network structures are regulated in the event of disease and the genes / proteins whose expression is altered are identified as the molecular target, studies will be carried out in accordance with this target. It has been found that different microbial agents or proteins / enzymes and metabolites they produce have the potential to prevent cancer or cause tumor shrinkage (Zitvogel et al., 2017). With this promising new microbiota-associated therapeutic approach, different preclinical studies have been conducted. One of these was the killing of cancer-associated bacteria using antibiotics. In a study conducted for this purpose, tumor volume shrinkage was observed as a result of trials performed on *Fusobacterium*-positive mouse xenografts of metronidazole (Bullman et al., 2017). However, it is thought that this approach is not an effective treatment strategy considering that most antibiotics do not have specific effect and that antibiotic use causes disbalance of the intestinal system (dysbiosis) or does not have the expected effect on correcting these disorders (Dethlefsen & Relman, 2011; Sullivan, 2001). Although the anti-cancer effects of probiotic ingestion of some beneficial bacteria have been reported (dos Reis et al., 2017) and there is insufficient data on clinical benefits, individual differences in colonization of different probiotic bacteria have been found to be high (Zmora et al., 2018). One of the clinical uses of the intestinal microbiota is fecal microbiome transfer (FMT), which is the application of fecal transfer from healthy person to diseased person. Although FMT has not yet been tested in patients with colon cancer, there have been trials with immunotherapy and chemotherapy (D'Haens & Jobin, 2019; McQuade et al., 2019). However, there is

a risk of contamination of different pathogens or diseases with FMT (Wong et al., 2019). In addition, small molecule inhibitors have been found to be effective in reducing the side effects of drugs and preventing cancer (Cougnoux et al., 2016). In short, instead of FMT applications, the development and transfer of microbial compositions with a high degree of stability are described as a more appropriate solution (Wirbel et al., 2019a).

3. MATERIALS AND METHODS

3.1. Public Data Collection

3.1.1. Public Data Collection of Whole Genome Shotgun Metagenomes

Raw data from nine publicly available fecal metagenomic sequencing (whole genome shotgun metagenome sequencing (WGS)) on CRC, adenomas and healthy controls were downloaded from the European Nucleotide Archive (ENA), using SRA-toolkit. ENA identifiers are: PRJEB27928 (Wirbel et al., 2019a), PRJNA447983 (Thomas et al., 2019), PRJEB6070 (Zeller et al., 2014), PRJEB10878 (Yu et al., 2017), PRJEB12449 (Vogtmann et al., 2016), PRJEB7774 (Feng et al., 2015), PRJDB4176 (Yachida et al., 2019), PRJDB4176 (Hannigan et al., 2018) and PRJNA531273 (Gupta et al., 2019).

3.1.2. Public Data Collection of 16S rRNA Amplicon Sequencing

16S rRNA sequencing raw data from eight publicly available studies (PRJEB6070 (Zeller et al., 2014), PRJNA290926 (Baxter et al., 2016), mothur (Zackular et al., 2014), PRJNA763023 (Yang et al., 2021), PRJEB36789 (Young et al., 2021), PRJNA698923 (Wang et al., 2021), PRJNA725613 (Liu et al., 2021) and PRJNA361327 (Hale et al., 2017) were also downloaded from ENA.

3.2. Data Preprocessing and Taxonomic Profiling

3.2.1. Data Preprocessing and Taxonomic Profiling for WGS Dataset

The metagenomic raw fastq files were quality controlled using FastQC tool (Andrews, 2010) and reads below 20 Phred scores and shorter than 50 bp were trimmed using the Trimmomatic tool (Bolger et al., 2014). Also, contaminated reads coming from human cells were removed by aligning metagenomics reads to hg19 decoy genome using bowtie2 aligner (Langmead & Salzberg, 2012). The Decoy genome includes the i) human reference genome, in addition to ii) sequences known as human reads from human whole genome-whole exome sequencing analyses, iii) sequences from other

known human genome assemblies, and iv) human DNA clones in bacteria (BAC and FOSMID). False positive results are eliminated by removing these contaminated reads that map to decoy genome (The Decoy Genome, n.d.). Clean reads obtained after quality trimming and removal of contaminated reads were used for taxonomic analysis. The MetaPhlAn2 (Metagenomic Phylogenetic Analysis) (Segata et al., 2012; Truong et al., 2015) pipeline was used to infer the taxonomic abundance profiles from CTR, CRC and ADN samples. MetaPhlAn2 implements a reference-based method for taxonomic profiling, in which metagenomic reads align with clade-specific marker genes to which known microorganisms are attached. The database used in MetaPhlAn2 analysis contains specific marker genes of ~13,500 bacterial and archaeal, ~3,500 viral and ~110 eukaryotic clades. As a result of the taxonomic profiling, species-level identification was achieved and the relative abundances of microorganisms present in the samples were obtained.

3.2.2. Data Preprocessing and Taxonomic Profiling for 16S Dataset

Amplicon sequencing samples were analysed using the QIIME 2 (Bolyen et al., 2019) (2020.2 version) metagenome pipeline. QIIME 2 data types, consisting of artifacts (.qza) and visualizations (.qzv), were utilized. Artifacts contain data and metadata, while visualization files present statistical tables, interactive visualizations, or static images. Metadata mapping files in tsv or txt format were created to include sample identifiers, barcode sequences, linker primer sequences, and other descriptive columns, validated using Keemei (Rideout et al., 2016). Data import into QIIME 2 was performed by specifying the input semantic type. For paired-end fastq files used as raw input data, the semantic type 'SampleData[PairedEndSequencesWithQuality]' was used to create QIIME 2 artifacts. Quality control and feature table construction were performed using the DADA2 pipeline (Callahan et al., 2016), which detects and corrects Illumina amplicon sequence data, filters phiX reads, and removes chimeric sequences. The denoise-paired method was used for paired-end data, resulting in an amplicon sequence variant (ASV) table.

Taxonomic analysis was conducted using the Silva database (SSU (small subunit-16S) 138 released in 2019) and QIIME2's naive Bayesian classifier built on Scikit-learn. The classifier was either trained on the exact region of the amplicon sequences or used a pre-trained dataset. The Silva 99% database trimmed to the V4 region (515F-806R

primers) was used, producing taxonomy files including counts and taxonomic bar plots. Subsequently, the count data were converted to relative abundances using *make_relative* function from *funrar* R package (version 1.5.0).

3.3. Enterotype Clustering and Classification

Enterotype (ET) analysis was performed using the genera abundance in each sample as described by Arumugam et al. (2011). R platform (version 4.1) was used for the analyses and visualizations. Genera with an average relative abundance of less than 1% were filtered out and normalized with square root transformation prior to clustering. CTR samples were clustered using probability distribution distance metric Jensen–Shannon divergence (JSD) and the Partitioning Around Medoids (PAM) clustering algorithm in the R package (version 4.1) “cluster”. The optimal number of clusters was determined using the Calinski-Harabasz (CH) Index (Calinski & Harabasz, 1974). Cluster validation was performed using the Silhouette method (Rousseeuw, 1984) and samples having zero silhouette coefficients (SH) were filtered out as the outliers. Clustering results were visualized with Principal-coordinate analysis (PCoA) in using the R package “ade4.”

It is known that diseases such as cancer are associated with intestinal dysbiosis, which represents a different gut microbiome profile compared to healthy people (Arumugam et al., 2011; Yang et al., 2019). Therefore, it is expected that patient samples would exhibit distinct microbiome profiles and would not cluster together with healthy samples into the same enterotypes. For this reason, the CRC and ADN samples were not subjected to a separate PAM clustering analysis. Instead, a K-Nearest Neighbors (KNN) approach was used to classify CRC and ADN samples into their respective enterotypes based on the robust microbial composition observed in the gut microbiome of healthy individuals.

In this classification process, the taxonomic profiles generated from CRC and ADN samples in the cohort were compared to the reference data of healthy individuals with the KNN algorithm, using the *knn* function in the R package “FNN”. The KNN algorithm is a supervised machine learning technique that classifies objects based on the similarity of their neighboring samples' classes, determined by their distance in feature space. In our case, the classification labels were the enterotype classes (ET1,

ET2, ET3), and the reference data consisted of the relative abundances of bacterial genera in the stool samples of healthy individuals.

To determine the number of nearest neighbors (k) for the KNN algorithm, the square root of the total number of samples in the CRC and ADN groups was taken. If the square root value was not an integer, the two closest integers were chosen as the k values, and the output with the best classification performance was selected. The Euclidean distance between the k nearest neighbors was calculated, and the class label of the nearest neighbors was assigned accordingly. Finally, the average of the class labels from the nearest neighbors was used to assign the enterotype class to each sample. To ensure high classification accuracy, samples with less than a 15% difference between the averages of their nearest neighbors were filtered out.

This approach allowed for the classification of CRC and ADN samples into their respective enterotypes based on the similarity of their microbial compositions to those of healthy individuals.

3.4. Confounder Analyses Based on Metadata

Recognizing the high risk of confounding in microbiome association studies, a comprehensive confounder analysis was conducted to examine the effects of potential confounding factors and disease status. For each enterotype class, the total variance of mean species abundances was compared to the variances explained by disease status (CTR vs. CRC, CTR vs. ADN) and confounding factors (age, BMI, sex, and study), resembling a linear model. Continuous values such as Age and BMI were transformed into discrete variables (as quartiles in Age factor, as groups representing lean (>25), overweight ($25-30$), and obese (>30) in the case of BMI) before variance calculations.

3.5. Identification of Enterotype Specific Microbial Species Using Differential Abundance Analysis (DAA)

The differential abundance analysis (DAA) was conducted to identify microbial markers specific to enterotypes associated within samples from each enterotype class were used for the DAA with Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC) (Lin & Peddada, 2020). ANCOM-BC is one of the differential abundance (DA) analysis methods which focused on correcting the bias induced by their differences among samples. DA of microbiome data is challenging

due the bias introduced by differences in the sampling fractions across samples. ANCOMBC R package (version 1.4.0) was used to implement ANCOM-BC analysis that is designed to determine taxa that are differentially abundant with respect to the covariate of interest. Study factor, which was found to be a confounding factor, was included as a covariate in the analyses. Significantly different species ($p < 0.05$) were identified in CRC and ADN cases compared to control samples (CTR). Additionally, a comparison between ADN and CRC was conducted to evaluate the differences in the adenoma-to-carcinoma sequence. Samples from various studies with fewer than five assignments to a group (CRC, CTR, or ADN) were excluded from the analysis.

3.6. Model Construction and Identification of ET-Specific Microbial Markers

Random Forest (RF) classification models were constructed using the SIAMCAT R package (version 1.14.0) (Wirbel et al., 2019a) to evaluate the importance of differentially abundant species and obtain robust microbial markers to distinguish cancer from control. First, relative abundances were filtered to eliminate markers with low overall abundance ($< 1E-03$). Subsequently, the data was log-transformed and standardized as z-scores. The dataset was then divided into training and test sets for 5 iterations of repeated 5-fold stratified cross-validation. In each iteration, a random forest model was trained on the training set and applied to predict the test set. Model performance was assessed by calculating the area under the ROC curve (AUC).

Various RF models were created with a specific emphasis on microbial markers that play a distinctive role in differentiating between enterotype groups, specifically within cancer, adenoma, and control samples. Differential abundances were used as features and binary classification between CRC-CTR, ADN-CTR and CRC-ADN samples were performed separately. The performance of the RF models was evaluated using three different feature sets: i) all abundance data (excluding enterotyping), ii) enterotype-specific abundance data, and iii) enterotype-specific differential abundance data.

3.7. Microbial Co-occurrence Analysis

Microbial co-occurrence analysis was also conducted to gain insights into the relationships between the human gut microbiome of CTR, CRC, and ADN samples. This analysis aimed to compare the co-occurrence patterns of microbial communities and investigate their associations within different enterotypes.

For the analysis, species-level microbial abundance profiles were used. To focus on more prevalent taxa, species with a total relative abundance of less than 1% were filtered out. Binary microbial interactions at the species level were determined using the "make network" function in the "phyloseq" package within the R platform. A maximum Bray-Curtis distance of 0.5 was employed as the threshold to define the strength of associations. Subsequently, network graphs were generated for each enterotype using the "igraph" R package (v.1.6.0) (Csárdi & Nepusz, 2006). In these network graphs, microbial taxa were represented as nodes, and the co-occurrence relationships were depicted as edges.

3.8. Functional Gene Family and Pathway Analysis of Enterotypes

Functional profiling was performed using the WGS dataset only, as it contains comprehensive genomic information necessary for accurate analysis of functional genes and metabolic pathways.

The HUMAnN2 (Abubucker et al., 2012) pipeline was used to define the microbial gene and pathway abundance profiles of each ET class. The HUMAnN2 pipeline utilizes the ChocoPhlAn pangenome database, which encompasses over 4,000 bacterial, archaeal and fungal pan-genomes, containing more than 18 million gene clusters. Additionally, it includes over 3,000 viral pan-genomes, encompassing more than 100,000 gene clusters (Franzosa et al., 2018). Metagenomic reads were quality trimmed and aligned to this database using the bowtie2 algorithm (Langmead & Salzberg, 2012), enabling the annotation of aligned genes. Remaining unannotated reads were realigned to the UniRef90 (The UniProt Reference Clusters) (Suzek et al., 2015) protein database (containing protein clusters with 90% sequence identity) with the Diamond (Buchfink et al., 2015) algorithm to obtain gene-based annotations. Following the annotation of aligned genes and gene families, relative gene abundances

were calculated. Furthermore, annotated genes were assigned to eggNOG and KEGG ortholog groups, allowing to determine the relative abundances of pathways associated with these gene families. Pathway information was obtained from the MetaCyc database.

Genes and pathways with relative abundances above 1% were subjected to DAA LEfSe analysis, using LEfSe (Linear discriminant analysis Effect Size) (Segata et al., 2011). LEfSe estimates the biological consistency and effect size of distinctive features (including microbial genes/pathways) and calculates Linear Discriminant Analysis (LDA) scores to distinguish between two or more biological classes. The LEfSe tool from the biobakery platform⁴⁶ was utilized for the analyses. The genes/proteins and pathways found to be associated with CRC/ADN cases with the highest effect size (LDA score >2) were listed and compared within ETs.

3.9. Correlation Analysis of Microbial Markers for CRC and ADN

Enterotype specific microbial species correlations between selected microbial markers were performed based on their relative abundances and variable (age and BMI) information. Spearman correlation analysis was performed to test whether there is a positive and/or negative relationship between microbial markers detected for CRC and ADN, and different age groups and obesity. The correlation coefficient (ρ) was calculated using the `corr.test` function in the "psych" package on the R platform. Statistically significant negative or positive relationships were determined among microbial markers and age, BMI groups.

3.10. Microbial Gene/Protein Targeted Drug Screening

The presence of drug molecules targeting genes predominantly found in enterotypes with high effect values related to CRC and ADN was investigated. To achieve this, the relevant genes were screened in DrugBank (Wishart et al., 2018), a comprehensive drug-target protein database. As no directly targeted drug molecules were found by searching these genes in the database, a BLAST analysis was performed to identify homologous genes. These homologous genes were then screened in the same database, and the identified targeted drug molecules were listed.

4. RESULTS

4.1. Datasets Characteristics in Meta-analysis

Metagenomic data from nine studies and amplicon sequencing data from eight studies were examined to assess the differential profiles of the human gut microbiome in patients with CRC and ADN compared to healthy controls. The primary objective was to identify specific biomarkers associated with adenoma and colon cancer. A total of 1579 whole metagenome sequencing (WGS) samples were collected, comprising 682 CRC patients, 207 ADN patients, and 690 healthy control subjects. A total of 1949 amplicon sequencing (16S) samples were collected, comprising 500 CRC patients, 485 ADN patients, and 964 healthy control subjects. **Table 4.1** and **Table 4.2** provide detailed information on the characteristics of the samples and demographic data.

Table 4.1: Characteristics of the fecal metagenomic datasets included in this study.

ENA identifier	Country	Sequencing Platform	Number of CTR samples	Number of ADN samples	Number of CRC samples	Age (average $\pm$ SD)	BMI (average $\pm$ SD)	Sex F(%) / M(%)
PRJEB27928	GER	Illumina HiSeq	60		52	60.7 $\pm$ 12.3	25.3 $\pm$ 3.4	42.9 / 57.1
PRJNA447983	ITA	Illumina HiSeq	51	23	61	67.4 $\pm$ 8.7	25.5 $\pm$ 3.9	35.0 / 65.0
PRJEB6070	FR	Illumina HiSeq	61	42	53	63.3 $\pm$ 10.9	25.3 $\pm$ 4.2	44.2 / 55.8
PRJEB10878	CHI	Illumina HiSeq	54		74	64.2 $\pm$ 9.1	23.8 $\pm$ 3.1	36.7 / 63.3
PRJEB12449	USA	Illumina HiSeq	51		52	61.5 $\pm$ 12.4	25.1 $\pm$ 4.2	29.1 / 70.9
PRJEB7774	AUS	Illumina HiSeq	63	47	46	66.9 $\pm$ 8.3	27.4 $\pm$ 4.0	43.7 / 56.4
PRJDB4176	JAP	Illumina HiSeq	292	67	286	61.7 $\pm$ 11.0	22.8 $\pm$ 3.3	42.6 / 57.4

Table 4.1: Continued.

ENA identifier	Country	Sequencing Platform	Number of CTR samples	Number of ADN samples	Number of CRC samples	Age (average $\pm$ SD)	BMI (average $\pm$ SD)	Sex F(%) / M(%)
PRJNA389927	USA	Illumina HiSeq	28	28	28	58.6 $\pm$ 10.7	28.1 $\pm$ 6.7	42.9 / 57.1
PRJNA531273	IND	Illumina NextSeq 500	30		30	50.6 $\pm$ 16.0	21.5 $\pm$ 3.0	50.0 / 48.3

GER: Germany, ITA: Italy, FR: France, USA: United States of America, CHI: China, AUS: Australia, JAP: Japan, IND: India.

4.2. Enterotype Identification

PAM clustering analysis was performed separately for WGS and 16S datasets on the CTR cohorts, which included 690 and 964 samples, respectively. The optimal number of clusters was determined to be 3 ($k=3$), and the samples were classified into three enterotypes based on their dominant genera.

Table 4.3 and **Table 4.4** present the total proportion of the relative abundances of the three most frequently detected genera across all CTR samples for WGS and 16S datasets, respectively. Additionally, mean and maximum relative abundances of these genus among samples are also listed in **Table 4.3** and **Table 4.4** in percentages.

Table 4.3: Characteristics of the fecal WGS datasets included in this study.

	Genus	Total (%)	Mean $\pm$ SD (%)	Max (%)
ET1	<i>Eubacterium</i>	13.82	13.81 $\pm$ 10.15	60.74
	<i>Ruminococcus</i>	12.03	12.02 $\pm$ 11.01	61.29
	<i>Bifidobacterium</i>	10.98	10.96 $\pm$ 14.07	70.78
ET2	<i>Bacteroides</i>	36.92	36.85 $\pm$ 15	84.31
	<i>Eubacterium</i>	6.66	6.64 $\pm$ 7.98	48.76
	<i>Parabacteroides</i>	5.03	5.02 $\pm$ 5.77	44.25
ET3	<i>Prevotella</i>	37.39	37.34 $\pm$ 19.94	95.29
	<i>Bacteroides</i>	9.64	9.62 $\pm$ 8.04	36.59
	<i>Eubacterium</i>	7.69	7.67 $\pm$ 6.89	40.16

Table 4.4: Characteristics of the fecal 16S datasets included in this study.

	Genus	Total (%)	Mean $\pm$ SD (%)	Max (%)
ET1	<i>Blautia</i>	16.52	15.13 $\pm$ 10.69	49.66
	<i>Bacteroides</i>	8.62	7.89 $\pm$ 6.79	28.38
	<i>Eubacterium</i>	6.54	5.99 $\pm$ 3.86	21.29
ET2	<i>Bacteroides</i>	38.38	31.04 $\pm$ 14.45	84.65
	<i>Faecalibacterium</i>	10.03	8.11 $\pm$ 6.81	51.63
	<i>Alistipes</i>	5.67	4.58 $\pm$ 3.9	22.81
ET3	<i>Prevotella</i>	32.12	28.05 $\pm$ 16.87	84.44
	<i>Bacteroides</i>	10.95	9.55 $\pm$ 6.86	34.70
	<i>Faecalibacterium</i>	10.32	9.01 $\pm$ 7.25	44.55

For WGS dataset, The Principal coordinate analysis (PCoA) plot in **Figure 4.1A** demonstrates clustering into three enterotypes that were dominated by the 3 genera. Enterotype 1 exhibited a higher proportion of *Eubacterium-Ruminococcus*, accounting for approximately 26% of all genera. Enterotype 2 was characterized by a greater prevalence of *Bacteroides*, comprising around 37% of all genera. Lastly, enterotype 3 displayed a higher proportion of *Prevotella*, constituting approximately 37% of all genera. These results are consistent with previous enterotype based studies, where 36% of all genera belonged to *Bacteroides*, 36% to *Prevotella* and 29% to *Ruminococcaceae* (Martinez-Medina et al., 2021).

For 16S dataset, PCoA plot in **Figure 4.1B** also demonstrates similar clusters with WGS dataset. Enterotype 1 exhibited a higher proportion (around 17 %) of *Blautia* genus, which is a newly introduced genus which mostly includes species previously described in the genus *Ruminococcus* (Liu et al., 2021). Approximately 23% of all genera were comprised of *Blautia- Eubacterium* genera. Enterotype 2 included 38% of *Bacteroides* genus while Enterotype 3 included %32 of *Prevotella* genus, which were determined as dominant genera in related enterotype group.

As previously reported (Arumugam et al., 2011), the healthy human gut microbiome is commonly classified into three clusters: *Bacteroides*, *Prevotella*, and *Eubacterium-Ruminococcus* genus from the *Eubacteriales* order.

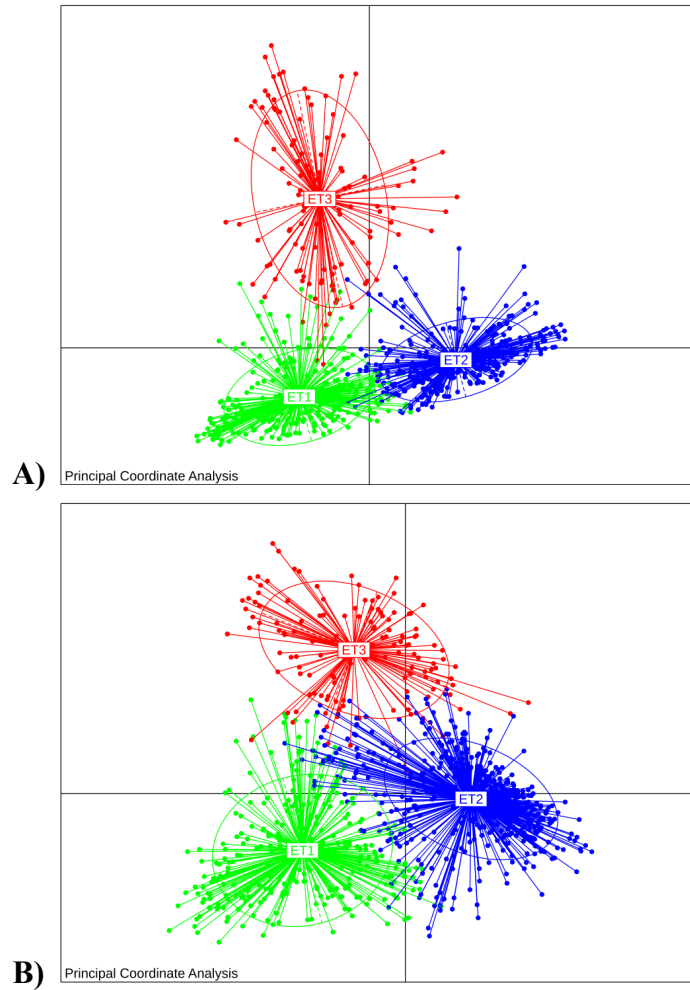


Figure 4.1: PCoA plot generated using the PAM clustering method based on Jensen-Shannon distance of microbial abundances for A) WGS dataset and B) 16S dataset. 1) *Eubacterium-Ruminococcus* (*Blautia*) dominated enterotype cluster (ET1), 2) *Bacteroides* dominated enterotype cluster (ET2), 3) *Prevotella* dominated enterotype cluster (ET3).

Table 4.5: Number of WGS samples assigned to each enterotype by clustering/classification analysis.

	<i>ET1</i>	<i>ET2</i>	<i>ET3</i>	<i>ET1</i> <i>filtered*</i>	<i>ET2</i> <i>filtered*</i>	<i>ET3</i> <i>filtered*</i>	% Filtered *
Number of CTR samples	281	293	116	251	289	110	5.8
Number of CRC samples	233	354	94	208	336	92	6.6
Number of ADN samples	100	79	26	92	70	24	9.3

* Samples exhibiting a difference of less than 15% among the average distance values of their K nearest neighbours with the corresponding class label were filtered

Table 4.6: Number of 16S samples assigned to each enterotype by clustering/classification analysis.

	<i>ET1</i>	<i>ET2</i>	<i>ET3</i>	<i>ET1 filtered*</i>	<i>ET2 filtered*</i>	<i>ET3 filtered*</i>	% Filtered *
Number of CTR samples	284	548	132	205	541	128	9.34
Number of CRC samples	201	238	61	174	225	57	8.8
Number of ADN samples	160	264	61	150	253	60	4.5

* Samples exhibiting a difference of less than 15% among the average distance values of their K nearest neighbours with the corresponding class label were filtered.

4.3. Identification of Potential Confounders

Microbial association studies are often at a high risk of confounding; thus confounder analysis was performed in order to determine potential confounding factors that may affect the differential analysis.

For all enterotype groups, the variance of species abundance explained by study factor was greater than the variance explained by disease status (Control vs. Cancer and Adenoma). **Figure 4.2** and **Figure 4.3** represent the variance plots obtained for WGS and 16S datasets, respectively. This underscores how the observed geographic variability in various studies (as depicted in **Table 4.1** and **Table 4.2**) affects fecal microbiome diversity, potentially limiting the generalizability of findings across diverse populations. Thus, the study factor, which played a pivotal role in contributing the highest level of observed variability in the dataset, was treated as a blocking factor in differential abundance analyses to identify microbial markers less affected by the study variable, contributing to more reliable and accurate microbial associations with adenoma and cancer.

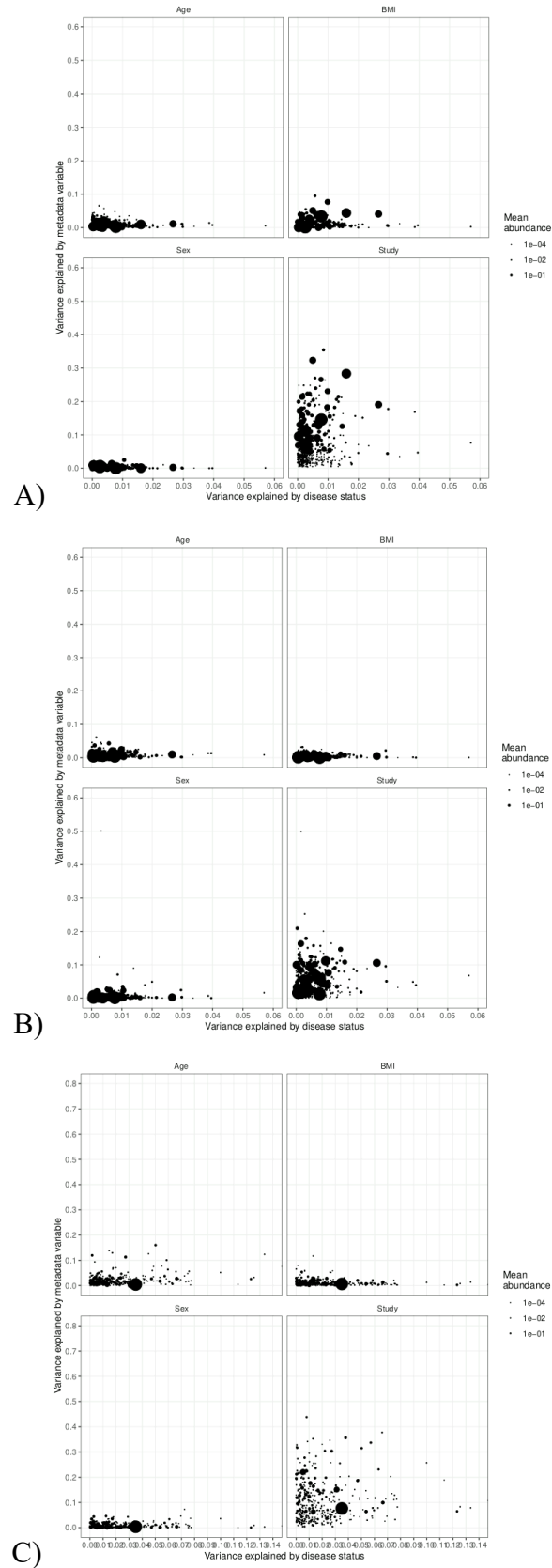


Figure 4.2: Variance explained by disease status (Control, CRC, ADN) in WGS dataset is plotted against variance explained by metadata variables (Age, BMI, Sex, Study) for microbial species in A) Enterotype 1, B) Enterotype 2 and C) Enterotype 3.

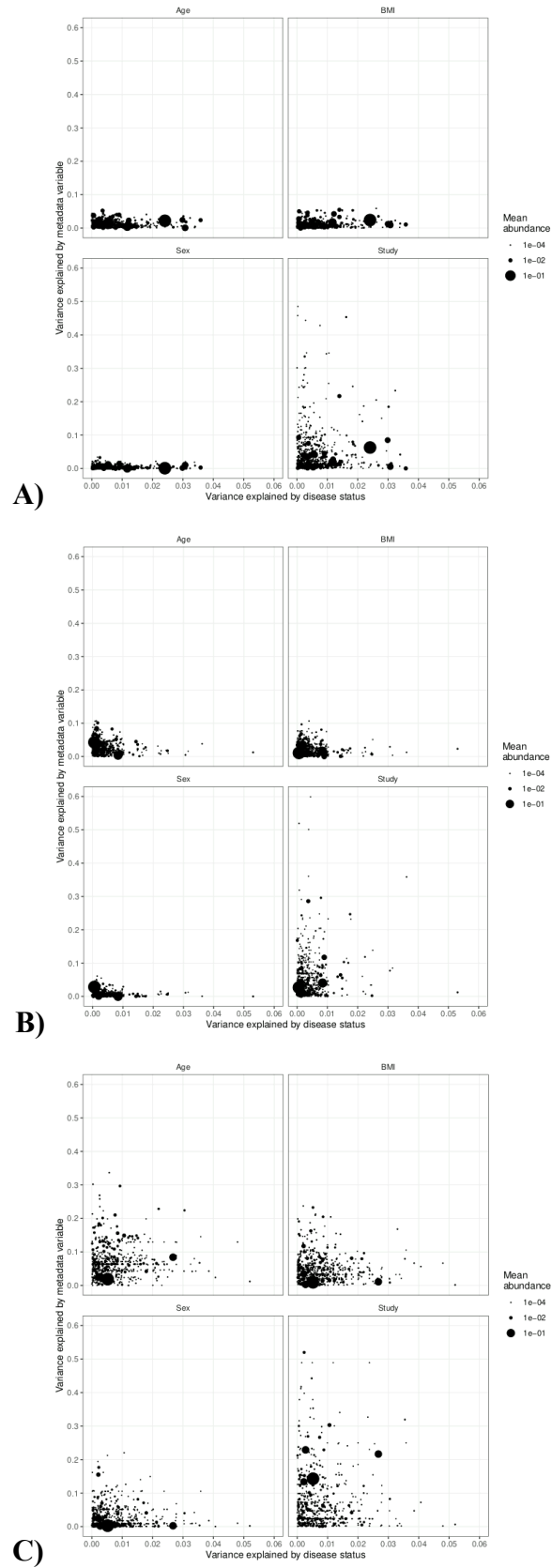


Figure 4.3: Variance explained by disease status (Control, CRC, ADN) in 16S dataset is plotted against variance explained by metadata variables (Age, BMI, Sex, Study) for microbial species in A) Enterotype 1, B) Enterotype 2 and C) Enterotype 3.

4.4. ET-Specific Differential Microbial Species

ET-specific DAA was performed using ANCOM-BC analysis with batch effect adjustments, leading to identification of differential microbial species less affected by the confounding study factor.

Following the ANCOM-BC analysis using WGS dataset, we observed significant differences ($p < 0.05$) in the abundance of 37 species (ET1), 48 species (ET2), and 32 species (ET3) between the healthy controls (CTR) and CRC samples. Similarly, significant differences were detected in 6 species (ET1), 15 species (ET2), and 13 species (ET3) between the CTR and adenoma (ADN) samples. Additional comparison between CRC and ADN samples was performed in order to identify differential species in the adenoma-carcinoma sequence. Significant differences were detected in 20 taxa (ET1), 31 taxa (ET2) and 14 taxa (ET3) between the ADN and CRC samples.

ANCOM-BC analysis using 16S dataset also revealed significantly changed ($p < 0.05$) microorganisms. No species level resolution was obtained for every differential microorganism, thus to make comparisons with WGS dataset, only significantly changed species were reported as follows: 27, 33, 13 species between CTR and CRC samples for ET1, ET2, ET3, respectively; 7, 1, 7 species between CTR and ADN samples for ET1, ET2, ET3, respectively and finally 13, 15, 8 species between CRC and ADN samples for ET1, ET2, ET3, respectively.

Upon comparing the results separately obtained for each enterotype group in each WGS and 16S datasets, we identified enterotype specific and non-specific differentially abundant species among the three enterotypes, as depicted in **Figure 4.4**, **Figure 4.5** and **Tables 4.7-4.12**. In WGS dataset, 17 CRC related microbial species were detected specific for ET1, 26 for ET2 and 15 for ET3, along with 9 common species detected for all enterotypes. In 16S dataset, 12 CRC related microbial species were detected specific for ET1, 21 for ET2, and 4 for ET3, along with 4 common species detected for all enterotypes. Among these separate ET comparisons detected using different sequencing methods (WGS or 16S), a few of them were found to be common. *Ruminococcus callidus*, *Bifidobacterium bifidum*, *Bacteroides salyersiae* were commonly detected as ET1 specific differential species using both dataset results. However, *B. bifidum* was found to be enriched in CTR samples using WGS dataset (p value 0.01) while it was enriched in CRC samples when 16S dataset used in the

analysis (p value 0.03). *Streptococcus mutans* was found to be commonly (in both 16S and WGS dataset) depleted in ADN when compared to CTR samples in ET1. *Prevotella copri* was commonly found to be depleted in CRC samples within ET3, along the adenoma-carcinoma sequence.

Using WGS dataset, the number of ADN related microbial species specific for ET1, ET2 and ET3 were found to be 5, 12 and 9, while 7, 7 and 1 numbers of differential microbial species were detected using 16S dataset, respectively. Interestingly, there were no markers that were common across all enterotypes in both WGS and 16S analyses, indicating distinct microbial signatures for each enterotype. On the other hand, the number of microbial species specific for ET1, ET2 and ET3 in adenoma-carcinoma sequence were detected as 11, 26 and 10 using WGS dataset and detected as 8, 12, 6, respectively. There were no markers that were common across all enterotypes in 16S dataset, while there were 2 common markers for adenoma-carcinoma sequence in all ETs using WGS dataset.

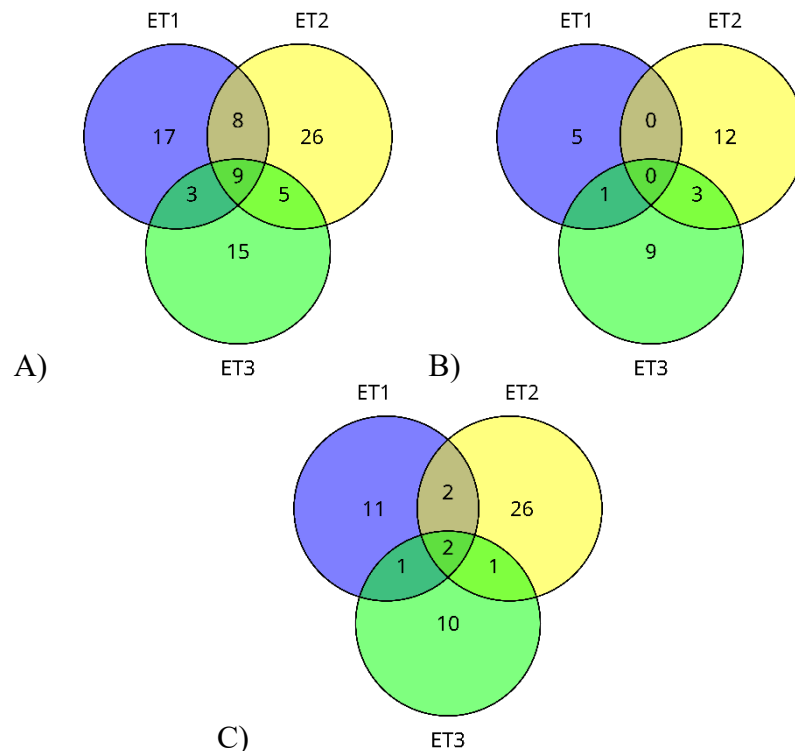


Figure 4.4: Comparison of differentially abundant microbial species for A) CRC-CTR, B) ADN-CTR and c) ADN-CRC comparison in each 3 enterotypes detected using WGS dataset.

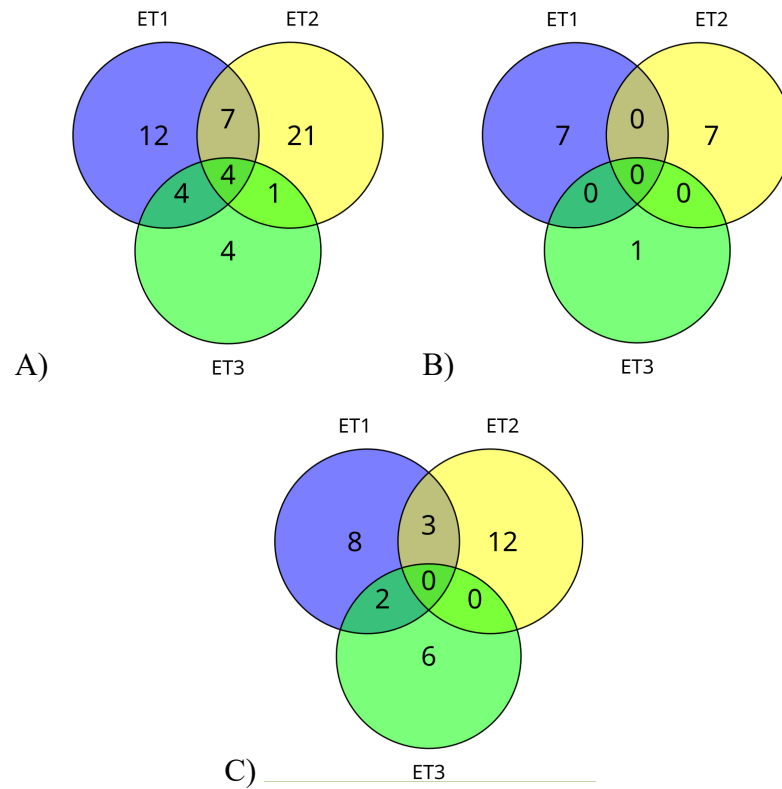


Figure 4.5: Comparison of differentially abundant microbial species for A) CRC-CTR, B) ADN-CTR and c) ADN-CRC comparison in each 3 enterotypes detected using 16S dataset.

Table 4.7: ET specific differential species related to CRC, detected from WGS dataset.

	Species Name	Enriched in	P-value
ET1	<i>Eubacterium rectale</i>	CTR	7.54×10^{-4}
	<i>Parabacteroides goldsteinii</i>	CRC	3.18×10^{-3}
	<i>Ruminococcus callidus</i> **	CTR	4.56×10^{-3}
	<i>Gordonibacter pamelaee</i>	CTR	5.17×10^{-3}
	<i>Streptococcus salivarius</i>	CTR	6.29×10^{-3}
	<i>Clostridium bolteae</i>	CRC	7.74×10^{-3}
	<i>Coprococcus comes</i>	CTR	1.04×10^{-2}
	<i>Bifidobacterium bifidum</i> **	CTR	1.19×10^{-2}
	<i>Bifidobacterium catenulatum</i>	CTR	1.38×10^{-2}
	<i>Bifidobacterium pseudocatenulatum</i>	CTR	1.56×10^{-2}
	<i>Bacteroides salyersiae</i> **	CRC	2.10×10^{-2}
	<i>Streptococcus anginosus</i>	CRC	2.52×10^{-2}
	<i>Lachnospiraceae bacterium 2158FAA</i>	CRC	2.70×10^{-2}
	<i>Desulfovibrio desulfuricans</i>	CRC	2.72×10^{-2}

Table 4.7: Continued.

	Species Name	Enriched in	P-value
ET1	<i>Bifidobacterium longum</i>	CTR	3.25x10 ⁻²
	<i>Ruminococcus sp 5 1 39BFAA</i>	CTR	3.40x10 ⁻²
	<i>Lachnospiraceae bacterium 3 1 46FAA</i>	CRC	3.74x10 ⁻²
ET2	<i>Sutterella parvirubra</i>	CTR	4.29x10 ⁻³
	<i>Granulicatella adiacens</i>	CRC	4.93x10 ⁻³
	<i>Morganella morganii</i>	CRC	5.87x10 ⁻³
	<i>Eubacterium bifforme</i>	CRC	9.21x10 ⁻³
	<i>Peptostreptococcus anaerobius</i>	CRC	1.02x10 ⁻²
	<i>Butyrivibrio crossotus</i>	CTR	1.22x10 ⁻²
	<i>Coprococcus catus</i>	CRC	1.42x10 ⁻²
	<i>Lactobacillus mucosae</i>	CRC	1.70x10 ⁻²
	<i>Prevotella buccae</i>	CRC	1.88x10 ⁻²
	<i>Flavonifractor plautii</i>	CRC	2.18x10 ⁻²
	<i>Bacteroides intestinalis</i>	CTR	2.26x10 ⁻²
	<i>Clostridium scindens</i>	CRC	2.39x10 ⁻²
	<i>Lactobacillus sanfranciscensis</i>	CRC	2.49x10 ⁻²
	<i>Megamonas hypermegale</i>	CTR	2.58x10 ⁻²
	<i>Alistipes sp APII</i>	CRC	2.70x10 ⁻²
ET3	<i>Veillonella atypica</i>	CRC	3.24x10 ⁻²
	<i>Anaerotruncus colihominis</i>	CRC	3.27x10 ⁻²
	<i>Coprobacter fastidiosus</i>	CTR	3.33x10 ⁻²
	<i>Holdemania filiformis</i>	CRC	3.59x10 ⁻²
	<i>Prevotella nigrescens</i>	CRC	3.81x10 ⁻²
	<i>Porphyromonas endodontalis</i>	CRC	3.82x10 ⁻²
	<i>Fusobacterium varium</i>	CRC	4.29x10 ⁻²
	<i>Eubacterium brachy</i>	CRC	4.63x10 ⁻²
	<i>Raoultella ornithinolytica</i>	CRC	4.72x10 ⁻²
	<i>Burkholderiales bacterium 1 1 47</i>	CTR	4.76x10 ⁻²
	<i>Clostridium sp ATCC BAA 442</i>	CRC	4.99x10 ⁻²
	<i>Veillonella parvula</i>	CRC	1.55x10 ⁻⁴
	<i>Bacteroidales bacterium ph8</i>	CTR	4.08x10 ⁻⁴
	<i>Dorea formicigenerans</i>	CRC	1.12x10 ⁻²
	<i>Ruminococcus gnavus</i>	CRC	1.46x10 ⁻²
	<i>Atopobium parvulum</i>	CRC	1.51x10 ⁻²
	<i>Bacteroides coprocola</i>	CTR	1.64x10 ⁻²

Table 4.7: Continued.

	Species Name	Enriched in	P-value
	<i>Alistipes senegalensis</i>	CTR	1.71x10 ⁻²
	<i>Lactobacillus ruminis</i>	CTR	1.78x10 ⁻²
	<i>Streptococcus mitis oralis pneumoniae</i>	CRC	2.41x10 ⁻²
	<i>Fusobacterium mortiferum</i>	CRC	2.54x10 ⁻²
	<i>Anaerostipes hadrus</i>	CTR	2.60x10 ⁻²
	<i>Clostridium nexile</i>	CRC	3.40x10 ⁻²
	<i>Bifidobacterium angulatum</i>	CTR	3.76x10 ⁻²
	<i>Alistipes putredinis</i>	CTR	4.39x10 ⁻²
	<i>Rothia mucilaginosa</i>	CRC	4.95x10 ⁻²
Common*	<i>Faecalibacterium prausnitzii</i>	CTR	8.29x10 ⁻⁶
	<i>Ruminococcus torques</i>	CRC	9.40x10 ⁻⁶
	<i>Peptostreptococcus stomatis</i>	CRC	1.48x10 ⁻⁴
	<i>Solobacterium moorei</i>	CRC	6.95x10 ⁻⁴
	<i>Bacteroides fragilis</i>	CRC	1.94x10 ⁻³
	<i>Fusobacterium nucleatum</i>	CRC	3.50x10 ⁻³
	<i>Parvimonas micra</i>	CRC	7.74x10 ⁻³
	<i>Bacteroides nordii</i>	CRC	3.75x10 ⁻²
	<i>Eubacterium eligens</i>	CTR	4.13x10 ⁻²

* Identified as markers in all enterotype groups (ET1, ET2 and ET3).

** Common with 16S dataset results.

Table 4.8: ET-specific differential species related to ADN when compared to CTR detected from WGS dataset.

	Species Name	Enriched in	P-value
ET1	<i>Streptococcus mutans</i> **	CTR	0.0027
	<i>Lactococcus lactis</i>	CTR	0.0146
	<i>Lachnospiraceae bacterium 7 1 58FAA</i>	ADN	0.0305
	<i>Bacteroides pectinophilus</i>	CTR	0.0447
	<i>Actinomyces sp ICM47</i>	ADN	0.0485
ET2	<i>Clostridium clostridioforme</i>	CTR	0.0012
	<i>Eubacterium sp 3 1 31</i>	CTR	0.0135
	<i>Streptococcus vestibularis</i>	CTR	0.0173
	<i>Lachnospiraceae bacterium 2 1 58FAA</i>	CTR	0.0231
	<i>Paraprevotella xylaniphila</i>	CTR	0.0236

Table 4.8: Continued.

	Species Name	Enriched in	P-value
ET2	<i>Bifidobacterium pseudocatenulatum</i>	CTR	0.0310
	<i>Oscillibacter</i> sp KLE 1745	CTR	0.0328
	<i>Coprococcus</i> sp ART55 1	CTR	0.0328
	<i>Eubacterium rectale</i>	ADN	0.0395
	<i>Enterobacter aerogenes</i>	ADN	0.0416
	<i>Lachnospiraceae bacterium</i> 8 1 57FAA	ADN	0.0472
	<i>Streptococcus anginosus</i>	CTR	0.0474
ET3	<i>Eubacterium ventriosum</i>	CTR	0.0003
	<i>Lachnospiraceae bacterium</i> 1 4 56FAA	CTR	0.0224
	<i>Alistipes onderdonkii</i>	CTR	0.0232
	<i>Bacteroides coprocola</i>	ADN	0.0361
	<i>Holdemania filiformis</i>	CTR	0.0394
	<i>Bacteroides thetaiotaomicron</i>	CTR	0.0406
	<i>Eubacterium hallii</i>	CTR	0.0409
	<i>Parabacteroides distasonis</i>	CTR	0.0434
	<i>Bacteroides stercoris</i>	CTR	0.0457

** Common with 16S dataset results.

Table 4.9: ET-specific differential species related to ADN when compared to CRC detected from WGS dataset.

	Species Name	Enriched in	P-value
ET1	<i>Gordonibacter pamelaiae</i>	ADN	0.0389
	<i>Peptostreptococcus anaerobius</i>	CRC	0.0262
	<i>Prevotella stercorea</i>	CRC	0.0495
	<i>Odoribacter splanchnicus</i>	CRC	0.0107
	<i>Porphyromonas asaccharolytica</i>	CRC	0.0289
	<i>Porphyromonas uenonis</i>	CRC	0.0194
	<i>Parabacteroides merdae</i>	CRC	0.0479
	<i>Faecalibacterium prausnitzii</i>	ADN	0.0384
	<i>Ruminococcaceae bacterium</i> D16	CRC	0.0449
	<i>Alistipes finegoldii</i>	CRC	0.0355
	<i>Escherichia coli</i>	CRC	0.0272

Table 4.9: Continued.

	Species Name	Enriched in	P-value
ET2	<i>Ruminococcus sp 5 1 39BFAA</i>	CRC	0.0150
	<i>Lachnospiraceae bacterium 7 1 58FAA</i>	CRC	0.0416
	<i>Streptococcus anginosus</i>	CRC	0.0006
	<i>Solobacterium moorei</i>	CRC	0.0006
	<i>Morganella morganii</i>	CRC	0.0469
	<i>Megasphaera elsdenii</i>	CRC	0.0266
	<i>Clostridium sp ATCC BAA 442</i>	CRC	0.0263
	<i>Streptococcus australis</i>	CRC	0.0446
	<i>Actinomyces odontolyticus</i>	CRC	0.0465
	<i>Parvimonas micra</i>	CRC	0.0009
	<i>Streptococcus vestibularis</i>	CRC	0.0045
	<i>Bacteroides plebeius</i>	CRC	0.0261
	<i>Lactococcus lactis</i>	CRC	0.0330
	<i>Granulicatella adiacens</i>	CRC	0.0043
	<i>Bifidobacterium pseudocatenulatum</i>	CRC	0.0334
	<i>Raoultella ornithinolytica</i>	CRC	0.0107
	<i>Lachnospiraceae bacterium 2 1 58FAA</i>	CRC	0.0092
	<i>Prevotella intermedia</i>	CRC	0.0211
	<i>Clostridium hathewayi</i>	CRC	0.0025
	<i>Lactobacillus casei paracasei</i>	CRC	0.0313
	<i>Coprococcus sp ART55 1</i>	CRC	0.0276
	<i>Streptococcus sanguinis</i>	CRC	0.0306
	<i>Eubacterium ramulus</i>	CRC	0.0420
	<i>Lachnospiraceae bacterium 3 1 57FAA CTI</i>	CRC	0.0107
	<i>Ruminococcus obeum</i>	CRC	0.0034
	<i>Alloprevotella tanneriae</i>	CRC	0.0272
ET3	<i>Streptococcus thermophilus</i>	CRC	0.0015
	<i>Prevotella copri**</i>	ADN	0.0457
	<i>Clostridium nexile</i>	CRC	0.0204
	<i>Eubacterium eligens</i>	ADN	0.0462
	<i>Ruminococcus gnavus</i>	CRC	2.02E-05
	<i>Eubacterium hallii</i>	CRC	0.0175
	<i>Phascolarctobacterium succinatutens</i>	ADN	0.0324
	<i>Lachnospiraceae bacterium 1 4 56FAA</i>	CRC	0.0188

Table 4.9: Continued.

	Species Name	Enriched in	P-value
ET3	<i>Parabacteroides distasonis</i>	CRC	0.0087
	<i>Ruminococcus torques</i>	CRC	0.0299
Common*	<i>Peptostreptococcus stomatis</i>	CRC	0.0038
	<i>Gemella morbillorum</i>	CRC	0.0133

* Identified as markers in all enterotype groups (ET1, ET2 and ET3).

** Common with 16S dataset results.

Table 4.10: ET-specific differential species related to CRC when compared to CTR detected from 16S dataset.

	Species Name	Enriched in	P-value
ET1	<i>Bifidobacterium bifidum</i> **	CRC	0.0344
	<i>Slackia piriformis</i>	CTR	0.0106
	<i>Bacteroides coprocola</i>	CTR	0.0313
	<i>Bacteroides salyersiae</i> **	CRC	0.0020
	<i>Alistipes sp.</i>	CRC	0.0343
	<i>Lactobacillus mucosae</i>	CTR	0.0495
	<i>Streptococcus sobrinus</i>	CRC	0.0172
	<i>Blautia faecis</i>	CRC	0.0402
	<i>[Clostridium] hylemonae</i>	CRC	0.0279
	<i>Ruminococcus callidus</i> **	CTR	0.0475
	<i>Ruminococcus champanellensis</i>	CTR	0.0262
	<i>Akkermansia muciniphila</i>	CRC	0.0497
ET2	<i>Bacteroides clarus</i>	CTR	0.0398
	<i>Bacteroides cutis</i>	CTR	0.0383
	<i>Bacteroides kribbi</i>	CRC	0.0165
	<i>Bacteroides nordii</i>	CRC	0.0067
	<i>Odoribacter splanchnicus</i>	CRC	0.0107
	<i>Sanguibacteroides justesenii</i>	CTR	0.0161
	<i>Prevotella intermedia</i>	CRC	0.0062
	<i>Prevotella stercorea</i>	CRC	0.0031
	<i>Alistipes finegoldii</i>	CRC	0.0219
	<i>Campylobacter showae</i>	CRC	0.0103
	<i>Campylobacter ureolyticus</i>	CRC	0.0056

Table 4.10: Continued.

	Species Name	Enriched in	P-value
ET2	<i>Desulfovibrio</i> sp.	CRC	0.0241
	<i>Firmicutes</i> bacterium	CTR	0.0094
	<i>bacterium</i> enrichment	CRC	0.0266
	<i>Clostridium aldenense</i>	CRC	0.0000
	<i>Clostridium scindens</i>	CRC	0.0007
	<i>Moryella indoligenes</i>	CRC	0.0272
	<i>Selenomonadales</i> bacterium	CRC	0.0472
	<i>Ruminococcus</i> sp.	CTR	0.0068
	<i>Ruminococcaceae</i> bacterium	CRC	0.0444
	<i>Dialister succinatiphilus</i>	CTR	0.0407
ET3	<i>Prevotella copri</i>	CTR	0.0106
	<i>Parabacteroides johnsonii</i>	CTR	0.0043
	<i>Massiliomicrobiota timonensis</i>	CTR	0.0467
	<i>Faecalitalea cylindroides</i>	CTR	0.0354
Common*	<i>Porphyromonas asaccharolytica</i>	CRC	0.0039
	<i>Porphyromonas somerae</i>	CRC	0.0118
	<i>Porphyromonas uenonis</i>	CRC	0.0308
	<i>Dialister pneumosintes</i>	CRC	0.0026

* Identified as markers in all enterotype groups (ET1, ET2 and ET3).

** Common with WGS dataset results.

Table 4.11: ET-specific differential species related to ADN when compared to CTR detected from 16S dataset.

	Species Name	Enriched in	P-value
ET1	<i>Actinomyces graevenitzi</i>	ADN	0.0239
	<i>Bacteroides eggerthii</i>	CTR	0.0097
	<i>Coprobacter secundus</i>	ADN	0.0300
	<i>Massiliprevotella massiliensis</i>	ADN	0.0342
	<i>Marseilla massiliensis</i>	CTR	0.0446
	<i>Lactobacillus mucosae</i>	CTR	0.0416
	<i>Streptococcus mutans</i> **	CTR	0.0468
ET2	<i>Bacteroides fluxus</i>	ADN	0.0121
	<i>Alistipes</i> sp.	CTR	0.0215

Table 4.11: Continued.

	Species Name	Enriched in	P-value
ET2	<i>Firmicutes bacterium</i>	CTR	0.0058
	<i>Eubacterium ramulus</i>	ADN	0.0128
	<i>Lachnospiraceae bacterium</i>	ADN	0.0326
	<i>Ruminococcus sp.</i>	CTR	0.0359
	<i>Veillonella atypica</i>	CTR	0.0033
ET3	<i>Parabacteroides merdae</i>	CTR	0.0039

** Common with WGS dataset results.

Table 4.12: ET-specific differential species related to ADN when compared to CRC detected from 16S dataset.

	Species Name	Enriched in	P-value
ET1	<i>Slackia isoflavoniconvertens</i>	CRC	0.0310
	<i>Massiliomicrobiota timonensis</i>	ADN	0.0327
	<i>Streptococcus mutans</i>	CRC	0.0454
	<i>Blautia hydrogenotrophica</i>	CRC	0.0133
	<i>[Clostridium] hylemonae</i>	CRC	0.0310
	<i>Mordavella sp.</i>	CRC	0.0430
	<i>Anaerotignum lactatifermentans</i>	CRC	0.0052
	<i>Clostridiales bacterium</i>	ADN	0.0396
ET2	<i>Bifidobacterium bifidum</i>	CRC	0.0409
	<i>Bacteroides coprophilus</i>	ADN	0.0391
	<i>Bacteroides fluxus</i>	ADN	0.0141
	<i>Gabonibacter massiliensis</i>	ADN	0.0426
	<i>Alistipes finegoldii</i>	CRC	0.0374
	<i>Parabacteroides gordonii</i>	ADN	0.0178
	<i>bacterium</i>	ADN	0.0019
	<i>Coprococcus eutactus</i>	ADN	0.0080
	<i>[Clostridium] scindens</i>	CRC	0.0020
	<i>Anaerotruncus sp.</i>	CRC	0.0348
	<i>Ruminococcus bicirculans</i>	ADN	0.0065
	<i>Burkholderiales bacterium</i>	ADN	0.0410
ET3	<i>Bacteroides plebeius</i>	ADN	0.0100
	<i>Prevotella copri**</i>	ADN	0.0350

Table 4.12: Continued.

	Species Name	Enriched in	P-value
ET3	<i>Parabacteroides johnsonii</i>	ADN	0.0233
	<i>Faecalitalea cylindroides</i>	ADN	0.0107
	<i>Megasphaera hexanoica</i>	ADN	0.0317
	<i>Parasutterella secunda</i>	ADN	0.0187

** Common with WGS dataset results.

4.5. Model Construction and Identification of ET-Specific Microbial Markers

5-fold cross-validation RF models were created by combining samples assigned to distinct enterotypes to assess the significance of each differentially abundant species in distinguishing cancer and adenoma from control. RF models, utilizing all species as features, exhibited comparable accuracy to models using only differentially abundant species. However, the latter approach, focusing on specific species, showed higher importance in classifying the top 10 most crucial marker species. RF models were constructed for both WGS and 16S dataset results.

For WGS dataset, the total mean contribution percentage of the top 10 important CRC markers to the models was observed as 46.8%, 45.7%, and 54.6% (**Figure 4.6**), while contribution levels were significantly lower when all features were employed to classify CRC and CTR samples. Specifically, the top 10 most important features in ET1, ET2, and ET3 groups were found to contribute 20%, 16.5%, and 19.6%, respectively. This underscores the enhanced discriminatory power achieved by focusing on differentially abundant species.

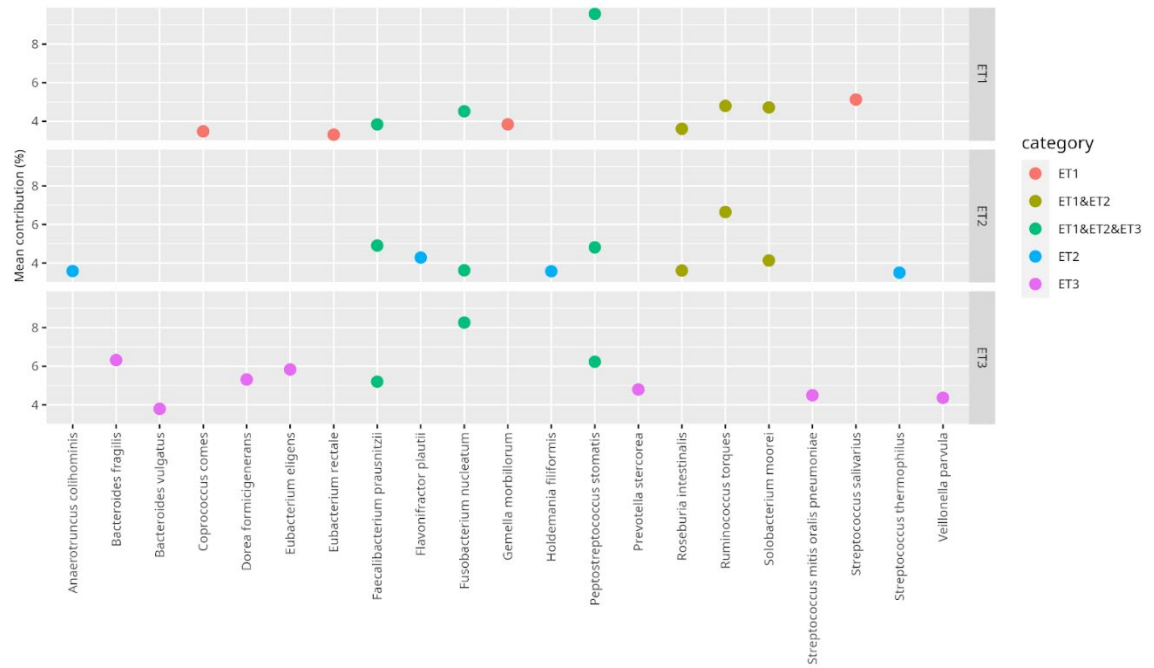
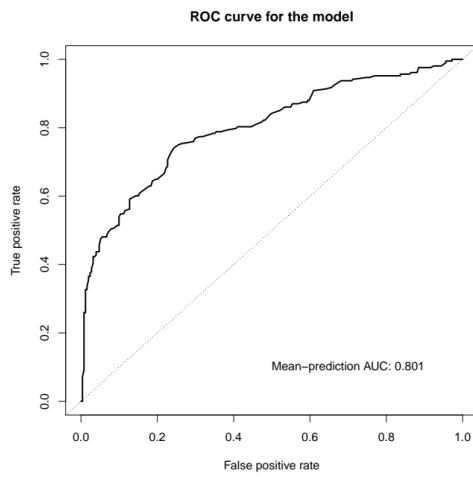
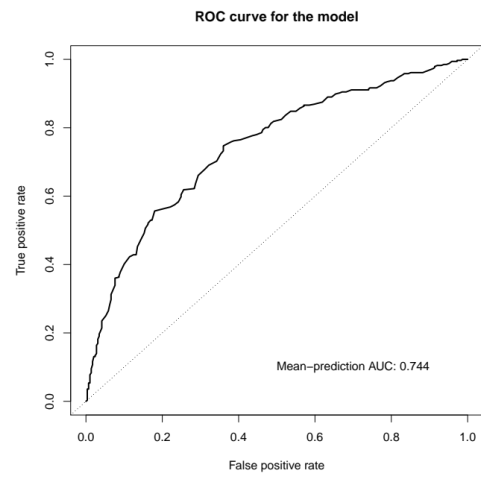


Figure 4.6: The mean contribution of top 10 marker species to the classification between CRC and CTR in each enterotype.

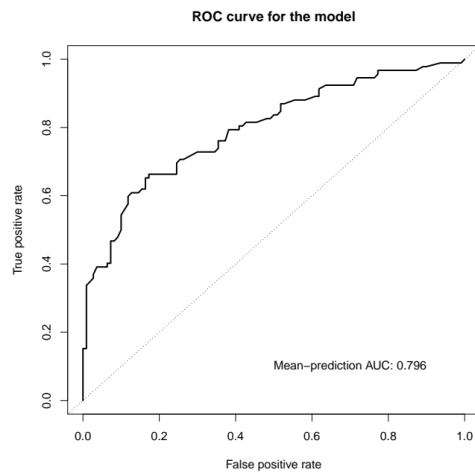
By utilizing differentially abundant species as features, sets of high-importance species were identified, comprising 34 ET1, 31 ET2, and 27 ET3 markers (account for more than 90% of the classifier weight) for distinguishing cancer from control, achieving AUCs of 0.80, 0.74, and 0.80, respectively (**Figure 4.7**). The distribution heatmaps of selected features across CRC and CTR samples are shown in **Figure 4.8**.



A)

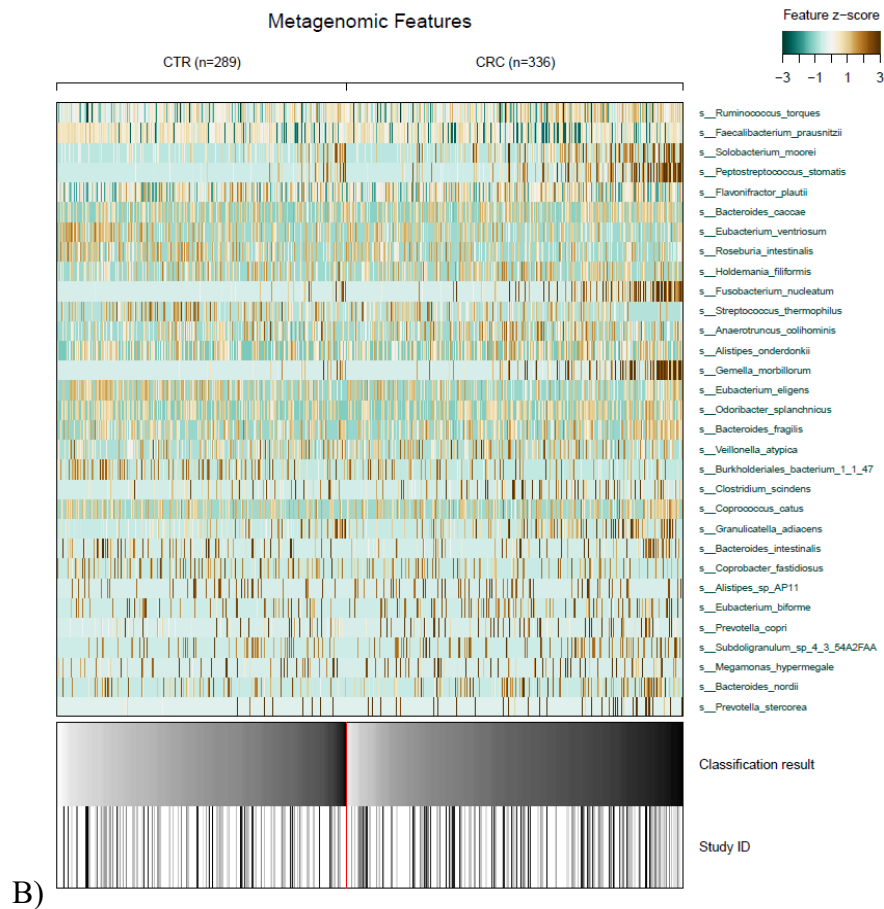
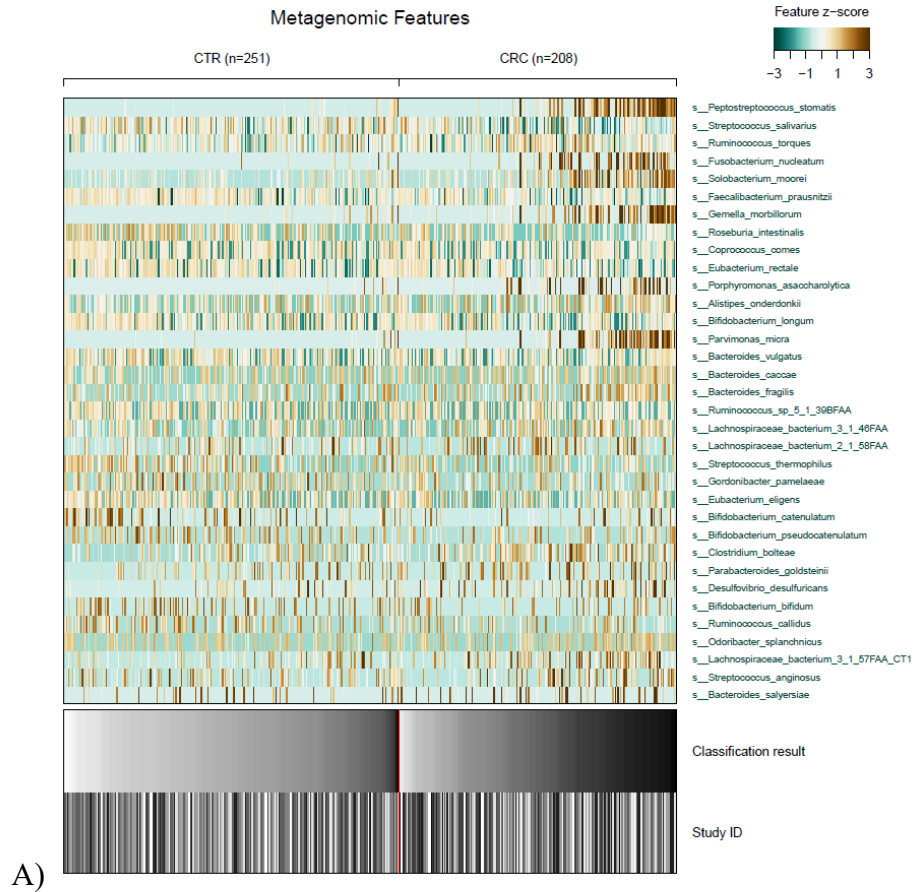


B)



C)

Figure 4.7: ROC curves of the Random Forest models for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.



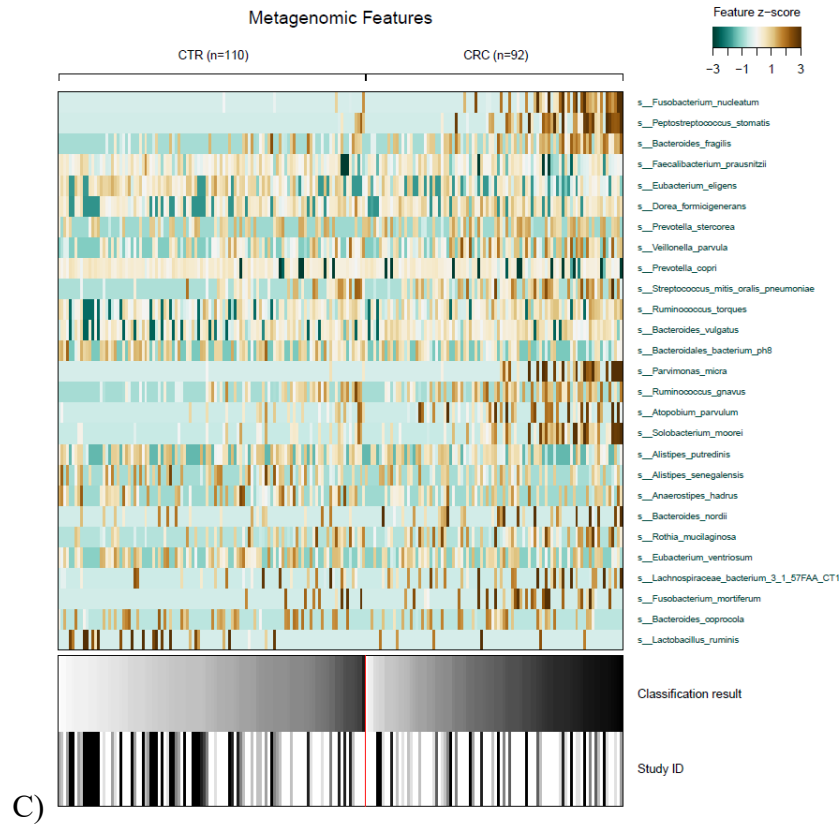


Figure 4.8: Heatmap visualizations of the Random Forest Classifier results for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.

Comparing the classification results between ET-nonspecific and ET-specific scenarios (**Figure 4.9**), we observed that 16 markers were specific to ET1, 14 to ET2, and 13 to ET3, with 6 markers common to all enterotypes. In the comprehensive analysis without considering enterotype labels, no enterotype-specific marker was identified, except for *Porphyromonas asaccharolytica* and *Streptococcus salivarius* in ET1.

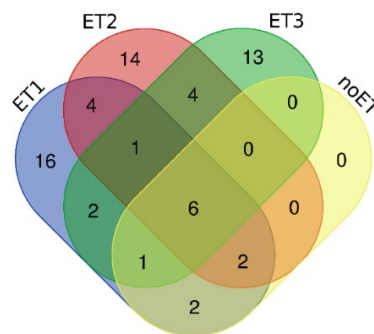


Figure 4.9: Venn diagram comparing Enterotype (ET) specific and non-ET specific CRC markers detected using WGS dataset.

For 16S dataset, RF models were constructed by utilizing differentially abundant species as features. The total mean contribution percentage of the top 10 important CRC markers to the models was observed as 74.2% (out of 27 markers), 87.1% (out of 33 markers) and 97% (out of 13 markers) for distinguishing cancer from control (Figure 4.10), achieving AUCs of 0.72, 0.73, and 0.66, respectively (Figure 4.11). The distribution heatmaps of selected features across CRC and CTR samples are shown in Figure 4.12.

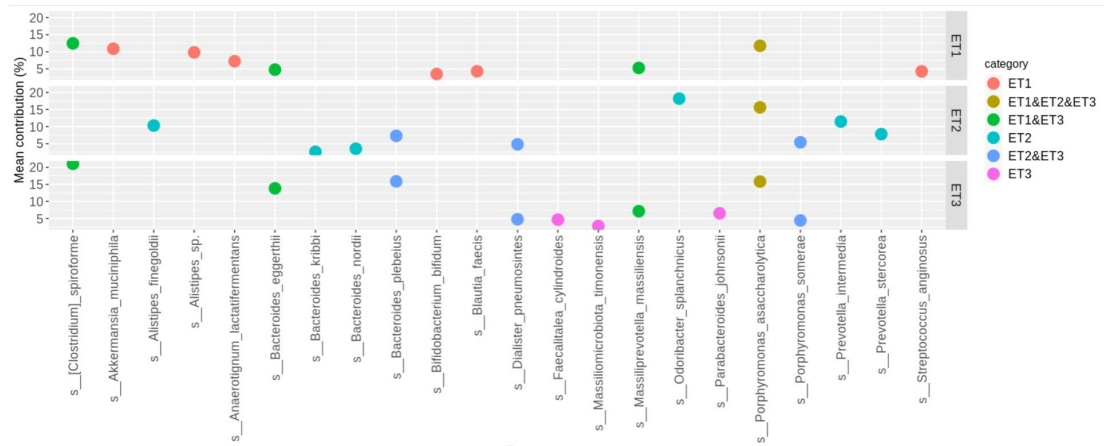


Figure 4.10: The mean contribution of top 10 marker species to the classification between CRC and CTR in each enterotype, using 16S dataset.

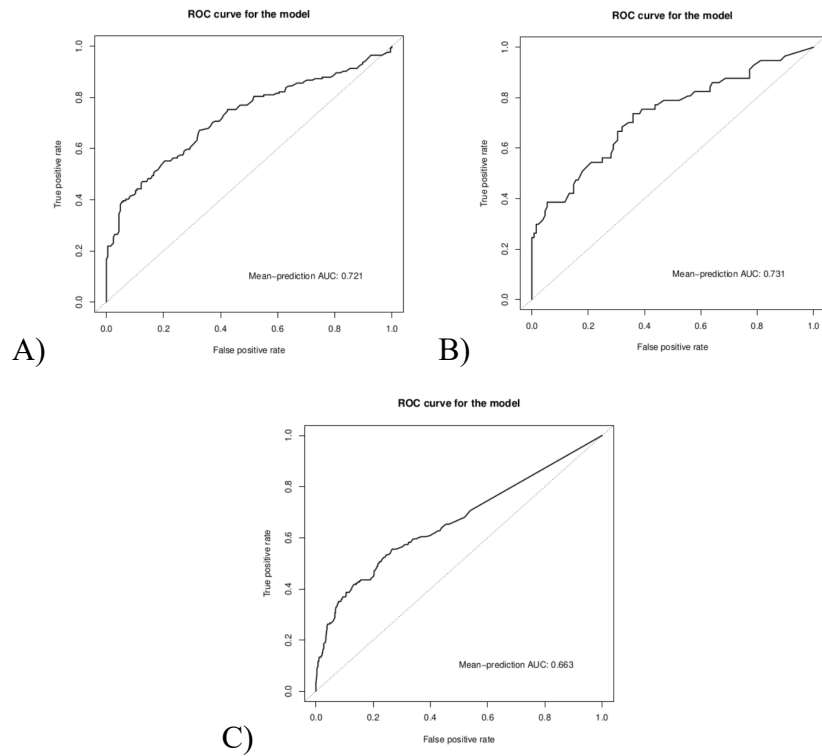
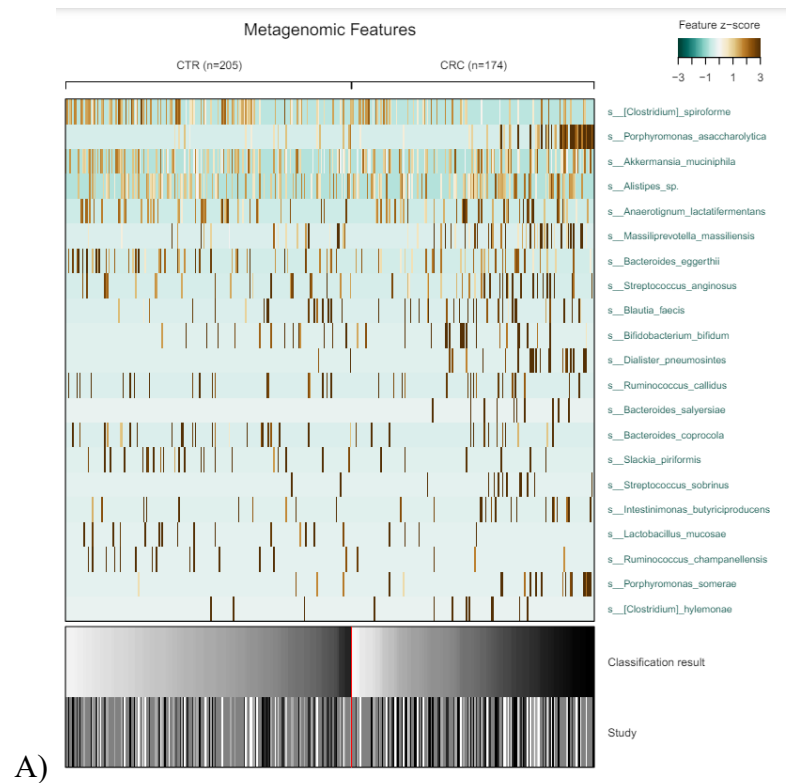


Figure 4.11: ROC curves of the Random Forest models for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.



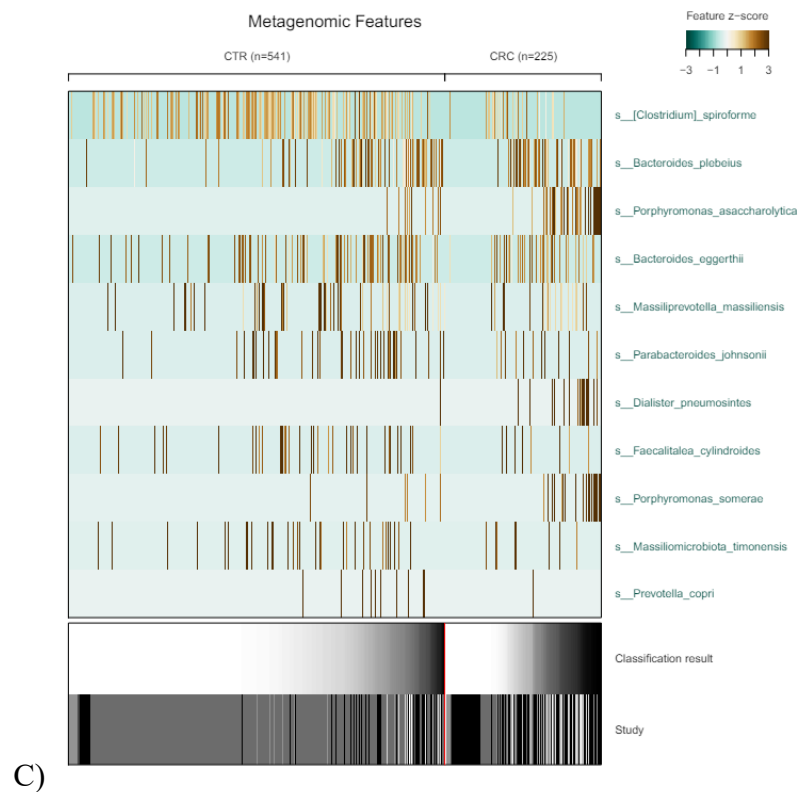
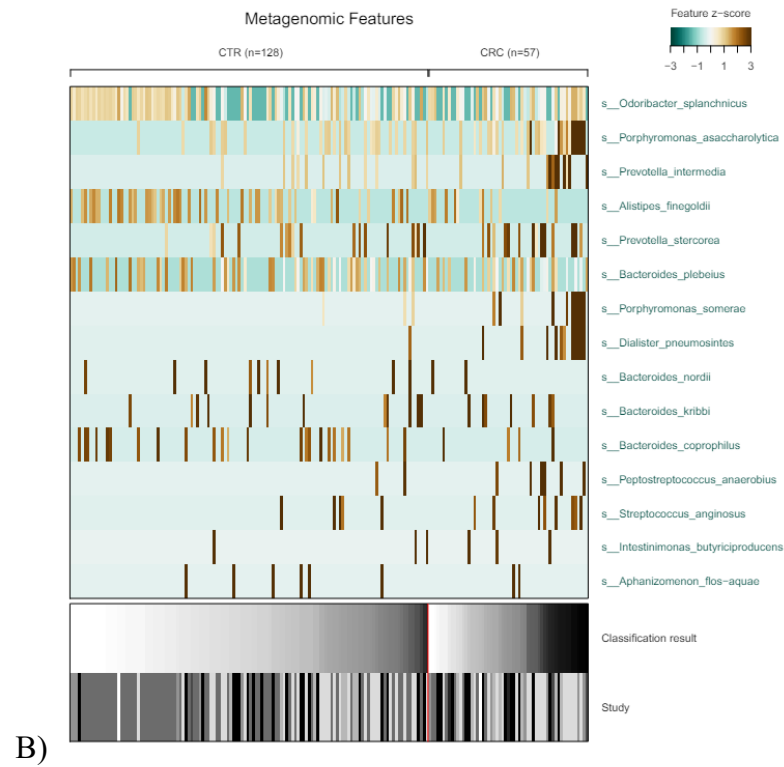


Figure 4.12: Heatmap visualizations of the Random Forest Classifier results for distinguishing CRC from Controls in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.

Predictive power of ADN associated markers was also evaluated by training RF models. Although specific species showed notable differences between ADN and CTR samples, the RF models struggled to distinguish ADN from controls with high accuracy for both 16S and WGS-based analysis. This was indicated by an area under the receiver operating characteristic (ROC) curve (AUC) of 0.63 (ET1), 0.61 (ET2), and 0.66 (ET3) for WGS-based models, while AUC of 0.61 (ET1), 0.5 (ET2), and 0.5 (ET3) were detected for 16-based models.

RF models to distinguish ADN from CRC states were also constructed using both WGS and 16S datasets separately. For WGS dataset, 11 ET1, 21 ET2, 10 ET3 markers (**Figure 4.13**) among differentially abundant species (account for more than 90 % of the classifier weight) were found to be discriminative between adenoma to carcinoma sequence with AUCs of 0.78, 0.7, 0.73, respectively (**Figure 4.14**). Top 10 most important markers are shown in **Figure 4.15** with total mean contribution percentages of 66 %, 50%, 80 % for ET1, ET2 and ET3, respectively. The distribution heatmaps of selected features across CRC and ADN samples are shown in **Figure 4.16**.

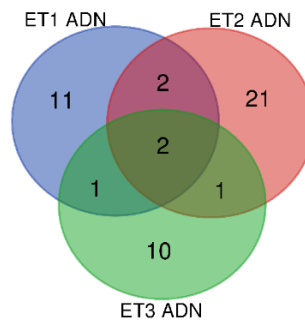


Figure 4.13: Venn diagram comparing ET specific ADN markers in the adenoma-carcinoma sequence.

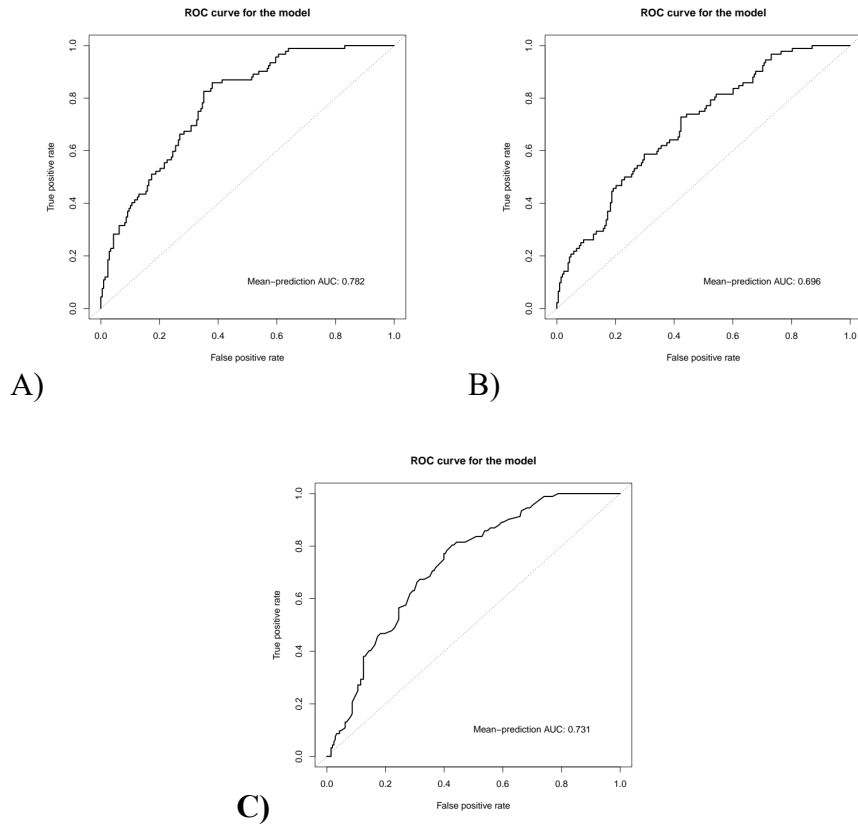


Figure 4.14: ROC curves of the Random Forest models for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.

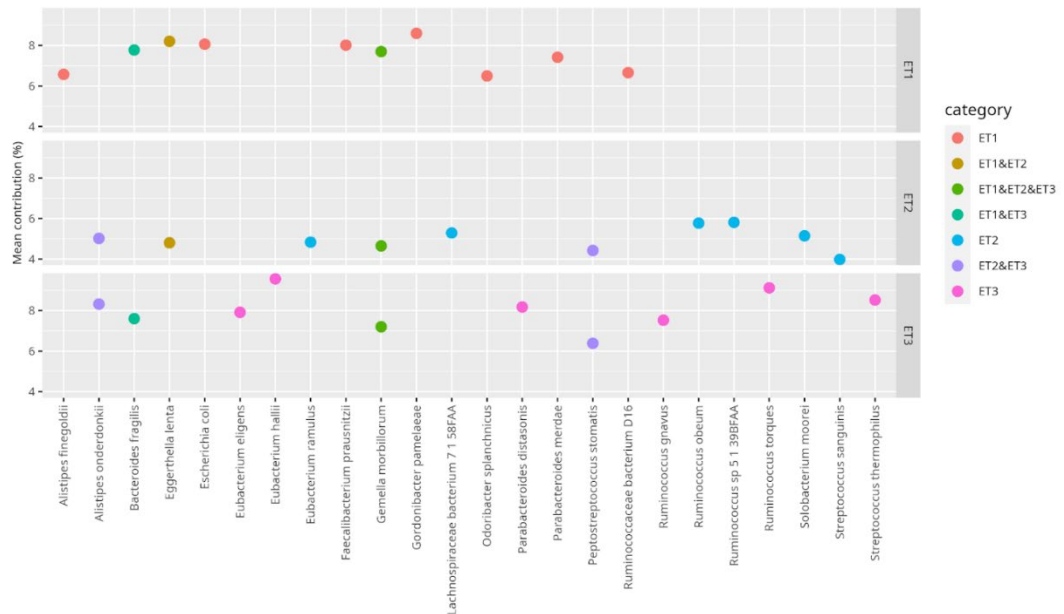
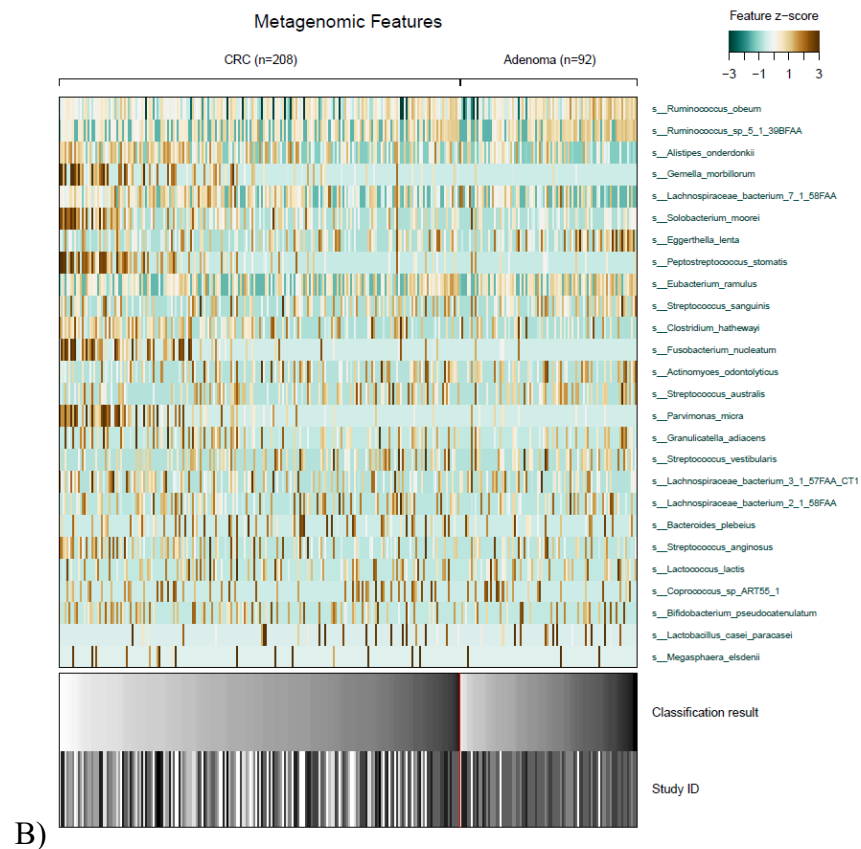
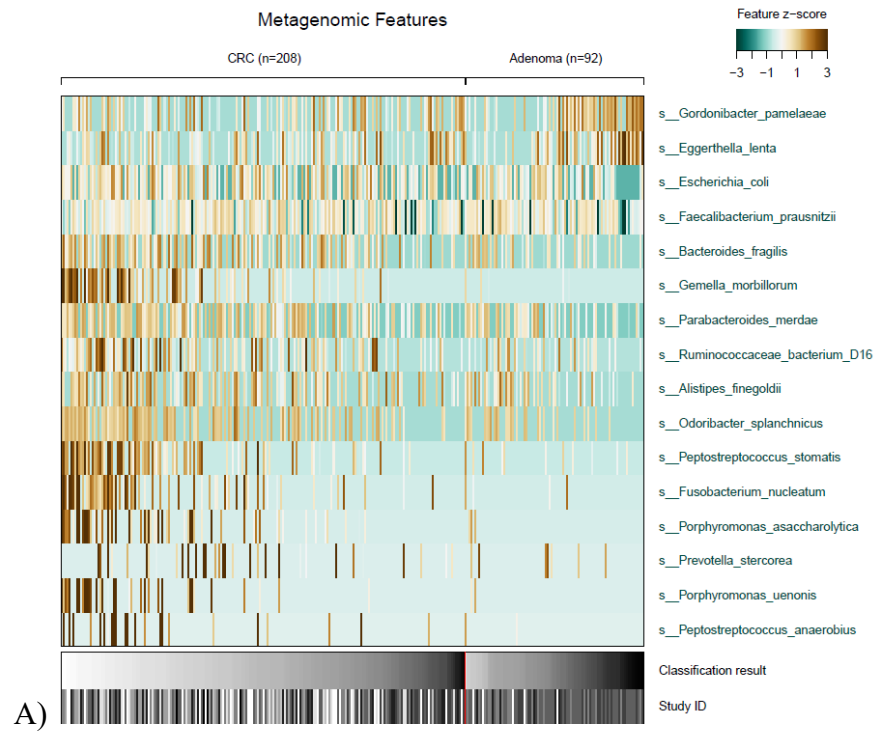


Figure 4.15: The mean contribution of top 10 marker species to the classification between CRC and ADN in each enterotype, using WGS dataset.



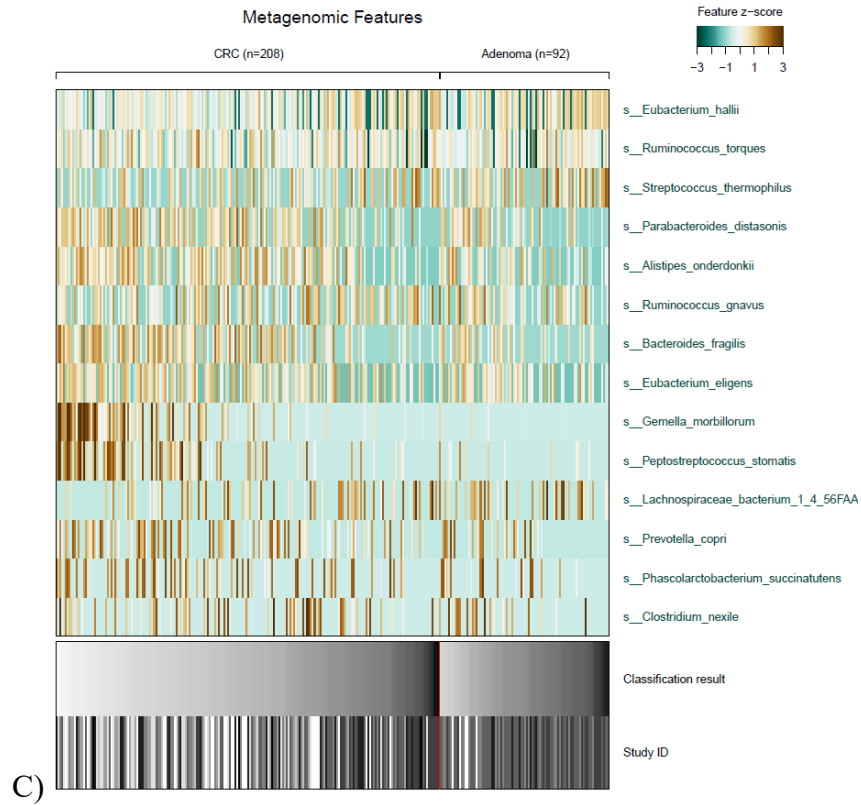


Figure 4.16: Heatmap visualizations of the Random Forest Classifier results for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using WGS dataset.

For 16S dataset, 12 ET1, 13 ET2, 8 ET3 markers among differentially abundant species were found to be discriminative between adenoma to carcinoma sequence with AUCs of 0.69, 0.66, 0.59 respectively (**Figure 4.17**). The top 10 most important markers are shown in **Figure 4.18** with total mean contribution percentages of 97.7 % (out of 13 markers), 98.2 % (out of 15 markers), 100 % (all 8 markers) for ET1, ET2 and ET3, respectively. The distribution heatmaps of selected features across CRC and ADN samples are shown in **Figure 4.19**.

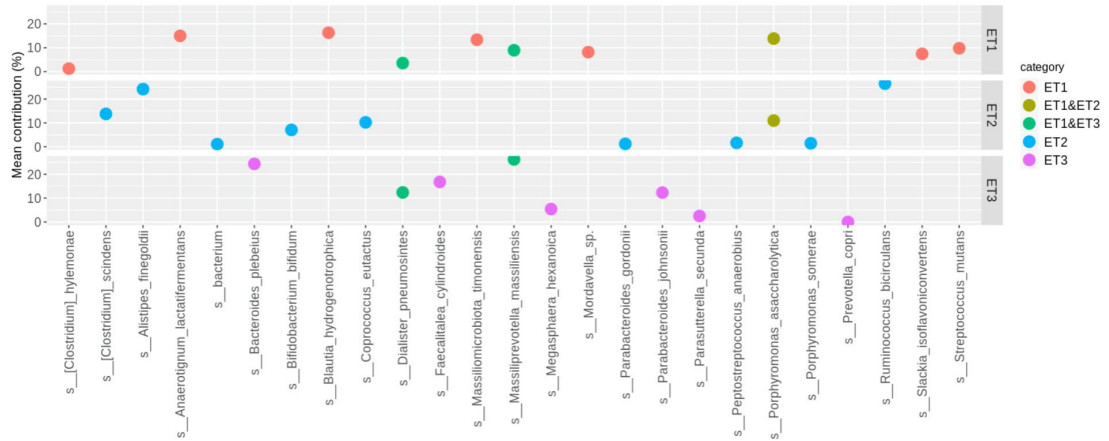


Figure 4.17: The mean contribution of top 10 marker species to the classification between CRC and ADN in each enterotype, using 16S dataset.

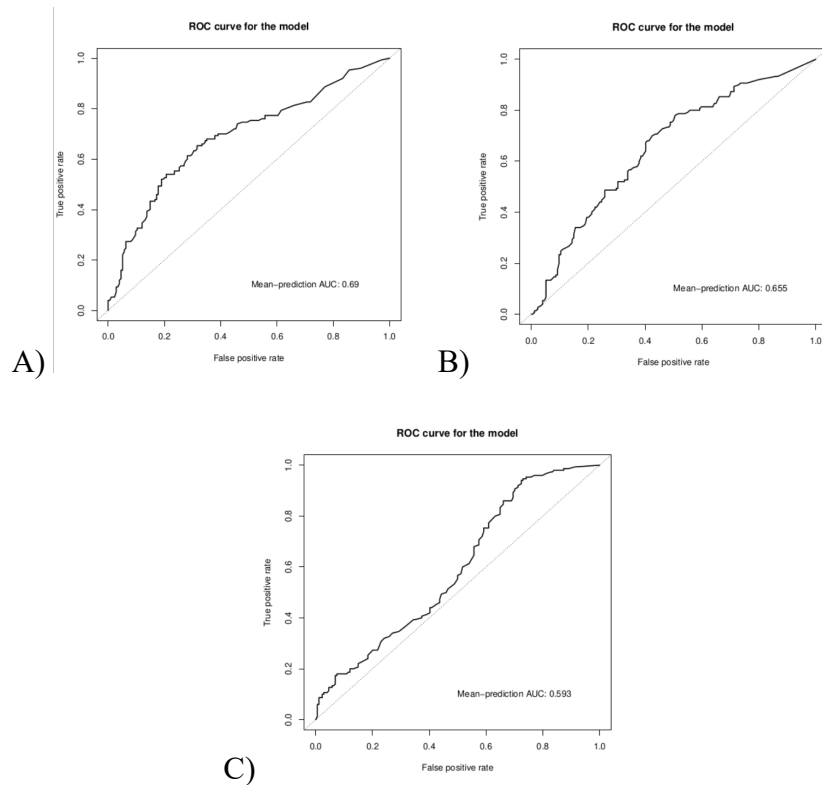
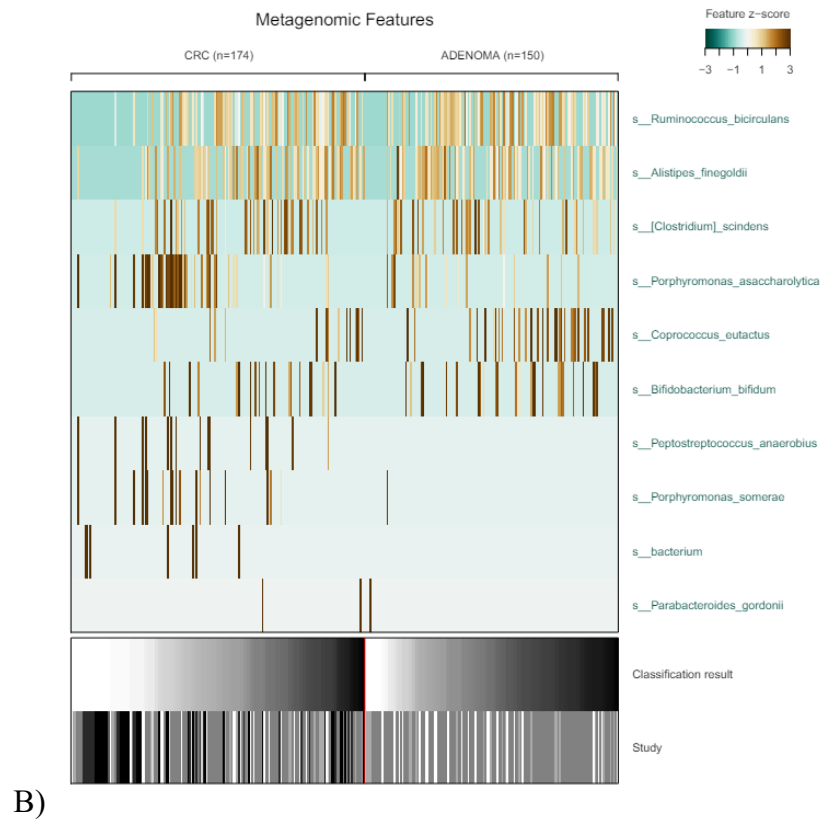
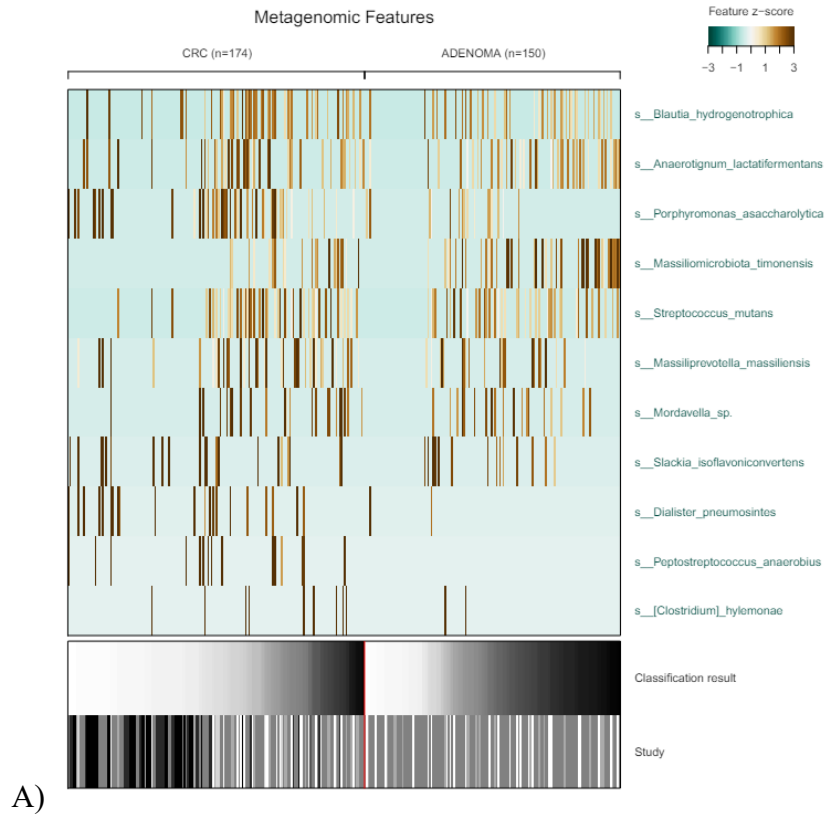


Figure 4.18. ROC curves of the Random Forest models for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.



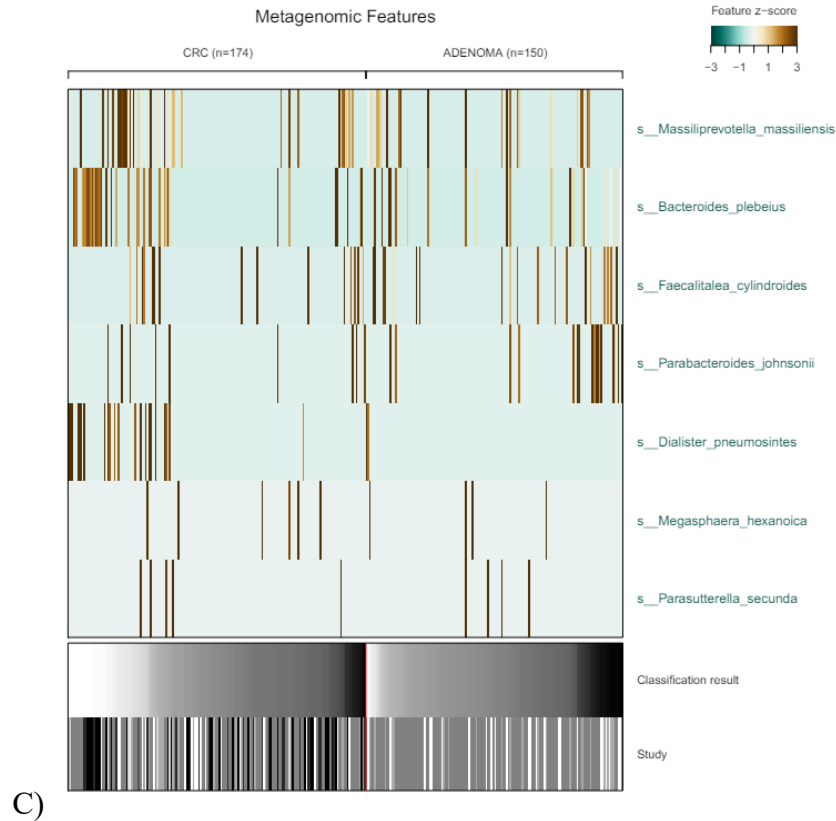


Figure 4.19: Heatmap visualizations of the Random Forest Classifier results for distinguishing ADN from CRC in A) Enterotype 1, B) Enterotype 2, C) Enterotype 3, detected using 16S dataset.

It is noteworthy that several microbial markers, including *Fusobacterium nucleatum*, *Ruminococcus torques*, *Bacteroides fragilis*, *Solobacterium moorei*, *Peptostreptococcus stomatis*, *Faecalibacterium prausnitzii* and *Eubacterium eligens* were found to be common CRC markers across all enterotypes. These markers were not identified as ET-specific differentiating features as they were consistently present regardless of the enterotype classification. These species have been widely studied as known CRC biomarkers in various research and meta-analyses (Baxter et al., 2016; Dai et al., 2018; Gupta et al., 2019; Ternes et al., 2020; Thomas et al., 2019; Wirbel et al., 2019b; Yachida et al., 2019; Yu et al., 2017; Zeller et al., 2014), regardless of the enterotype. *Faecalibacterium prausnitzii* and *Eubacterium eligens* markers exhibited differential enrichment in control samples compared to CRC. These probiotics, known for their effectiveness in CRC management and their role as crucial intestinal microbes in butyrate production, may contribute to the prevention of CRC (Avuthu & Guda, 2022; Dikeocha et al., 2022; Kaźmierczak-Siedlecka et al., 2020; Mizutani et al., 2020).

For differential species between adenoma-carcinoma sequence, *P. stomatis* and *Gemella morbillorum* species were found to be significantly enriched in carcinoma. *P. stomatis* was also detected as a common CRC marker in RF models as well, listed in the top 10 most important features. This is consistent with the studies that validate *P. stomatis* to be an early diagnosing marker for colon cancer (Shen et al., 2021; Ternes et al., 2020). *G. morbillorum* was detected to be differentially enriched in CRC samples with a high importance (acc. to RF model) only in adenoma-carcinoma sequence. This is also consistent with the studies looking for early diagnostic markers, where *G. morbillorum* was found to be amongst the most significantly abundant bacteria in early-stage CRC (Olovo et al., 2021).

4.6. Enterotype-Specific Co-occurrence Networks

The microbial co-occurrence analysis provided insights into the relationships and patterns of microbial communities within different enterotypes associated with CRC and ADN. By examining the co-occurrence of microbial species, potential associations relevant to CRC and ADN were explored.

The binary associations of microorganisms were studied with Bray-Curtis similarity based on relative abundance amounts in samples. Species relationships were identified within each ET, and a total of 20 binary associations were observed in ET1, 7 associations in ET2, and 24 associations in ET3 (**Figure 4.20**).

Enterotype 1 Network



Enterotype 2 Network



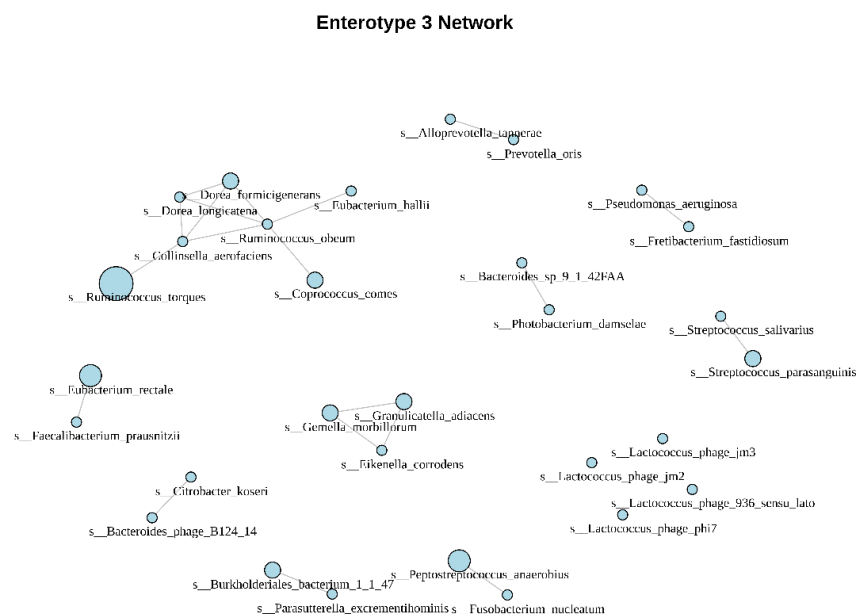


Figure 4.20. Microbial co-occurrence network plots for Enterotypes detected using WGS dataset.

Among these associations, the presence of the association between *Burkholderiales bacterium 1_1_47* and *Parasutterella excrementihomins* was observed in all three ETs. Additionally, common associations between *Eubacterium hallii* - *Ruminococcus obeum*, *Eubacterium rectale* - *Faecalibacterium prausnitzii*, *Ruminococcus obeum* - (*Coprococcus comes*, *Dorea formicigenerans*, *Dorea longicatena*) were found between ET1 and ET2, while the association between *Streptococcus parasanguinis* and *Streptococcus saliva* was shared between ET2 and ET3. *Solobacterium moorei* and *Parvimonas micra*, which were defined as common differentially abundant species in all ETs, were observed to be closely related, consistent with the literature reporting a strong relationship between CRC markers *P. micra* and *S. moorei* (Chung et al., 2018). Among the ET1-specific markers, *Coprococcus comes* was observed to be co-abundant with *Ruminococcus torques*, a common marker for all ETs. In ET2, a network characterized by weak microbial associations was detected using a Bray-Curtis distance threshold of 0.5, comprising only 7 distinct binary associations. Furthermore, additional network analyses were conducted to identify and compare co-abundance patterns in individual CTR, CRC, and ADN groups within each enterotype, and the resulting co-abundant microbial species detected in CRC, CTR, and ADN samples classified into three ETs are listed in **Table 4.13**.

Table 4.13: Co-abundant microbial species detected in CRC, CTR and ADN samples classified into three ETs.

Co-abundant species detected only in CTR samples		
ET1	<i>Clostridium bolteae</i> *	<i>Clostridium hathewayi</i>
	<i>Coprococcus comes</i> *	<i>Dorea longicatena</i>
	<i>Eubacterium hallii</i>	<i>Ruminococcus torques</i> ***
	<i>Porphyromonas asaccharolytica</i> **	<i>Peptostreptococcus stomatis</i> ***
	<i>Ruminococcus obeum</i>	<i>Ruminococcus torques</i> ***
	<i>Ruminococcus obeum</i>	<i>Coprococcus comes</i> *
	<i>Ruminococcus torques</i> ***	<i>Coprococcus comes</i> *
	<i>Streptococcus parasanguinis</i>	<i>Streptococcus salivarius</i> *
ET2	<i>Granulicatella adiacens</i> *	<i>Solobacterium moorei</i> ***
ET3	<i>Ruminococcus torques</i> ***	<i>Coprococcus comes</i>
Co-abundant species detected only in CRC samples		
ET1	<i>Gemella morbillorum</i> **	<i>Solobacterium moorei</i> ***
	<i>Parvimonas micra</i> ***	<i>Solobacterium moorei</i> ***
	<i>Eubacterium hallii</i>	<i>Ruminococcus sp 5 1 39BFAA</i> *
	<i>Collinsella aerofaciens</i>	<i>Ruminococcus obeum</i>
	<i>Collinsella aerofaciens</i>	<i>Dorea longicatena</i>
	<i>Clostridium hathewayi</i>	<i>Acidaminococcus sp D21</i>
ET2	<i>Porphyromonas asaccharolytica</i> **	<i>Porphyromonas uenonis</i>
ET3	<i>Collinsella aerofaciens</i>	<i>Ruminococcus torques</i> ***
	<i>Peptostreptococcus anaerobius</i>	<i>Fusobacterium nucleatum</i> ***
	<i>Rothia mucilaginosa</i> *	<i>Streptococcus mitis oralis pneumoniae</i> *
	<i>Gemella morbillorum</i> **	<i>Granulicatella adiacens</i>
	<i>Gemella morbillorum</i> **	<i>Eikenella corrodens</i>
Co-abundant species detected only in ADN samples		
ET1	<i>Clostridiaceae bacterium JC118</i>	<i>Clostridiales bacterium 1 7 47FAA</i>
	<i>Clostridium asparagiforme</i>	<i>Clostridium symbiosum</i>
	<i>Clostridiales bacterium 1 7 47FAA</i>	<i>Erysipelotrichaceae bacterium 2 2 44A</i>
	<i>Ruminococcus torques</i> ***	<i>Dorea longicatena</i>
ET3	<i>Methanobrevibacter smithii</i>	<i>Ruminococcus torques</i> ***
	<i>Collinsella aerofaciens</i>	<i>Ruminococcus torques</i> ***
	<i>Ruminococcus obeum</i>	<i>Ruminococcus torques</i> ***
	<i>Streptococcus parasanguinis</i>	<i>Veillonella parvula</i>
	<i>Ruminococcus torques</i> ***	<i>Ruminococcus sp 5 1 39BFAA</i>
	<i>Ruminococcus torques</i> ***	<i>Eubacterium bifforme</i>
	<i>Ruminococcus torques</i> ***	<i>Dorea formicigenerans</i>
	<i>Ruminococcus torques</i> ***	<i>Dorea longicatena</i>
	<i>Ruminococcus torques</i> ***	<i>Catenibacterium mitsuokai</i>
	<i>Veillonella parvula</i>	<i>Enterobacter aerogenes</i> *

* ET-specific marker, ** common CRC marker for two of the ETs, *** common CRC marker for all ETs.

When we evaluate the co-occurrence associations in individual groups, specific species were found to be co-abundant with ET-specific markers. *Coprococcus comes* was observed to have association with a high degree of connectivity, which was also detected as in top 10 most important ET1 specific markers according to our RF classifier (**Figure 4.6**). This species is known to be associated with a healthy gut, and depleted levels of *Coprococcus* have been observed in CRC patients (Flemer et al., 2017; W. Liu et al., 2016). In our analysis focused on ET1, *Coprococcus comes* was recognized as a marker depleted in CRC, indicating that the association with this species might have a crucial impact on the pathogenesis of CRC within this particular enterotype. Similarly, the common markers *Solobacterium moorei*, and *Parvimonas micra*, which are CRC-enriched species, showed increased abundance in CRC samples and were not observed in the microbial network of the CTR group. While no microbial co-occurrence association was detected among the specific markers in ET2, an interaction between the CRC-enriched marker *Porphyromonas asaccharolytica* and *Porphyromonas uenonis* was observed. This interaction has been reported in CRC samples but not in CTR samples in another study (Jiang et al., 2021). Regarding the ET3-specific CRC markers, *Rothia mucilaginosa* and *Streptococcus mitis oralis pneumoniae* (detected in top 10 most important marker as depicted in **Figure 4.6**), their coexistence with similar relative abundances suggests the potential for their collaborative role in contributing to disease progression in this specific enterotype. Co-abundance profiles revealed distinct associations in microbial compositions between disease and control conditions. Notably, the specific associations related to the selected markers can provide insights into the species jointly involved in cancer pathways. It is necessary to confirm these associated mechanisms can be confirmed with possible metabolite analyses.

4.7. Functional Gene Families and Pathways Associated with ET-Specific Microbial Markers

The functional gene family and pathway analysis identified potential functional signatures related to CRC and ADN within specific enterotypes. It is known that gut microbiota influences human metabolism, and different enterotype clusters may exhibit functional differences in genes and pathways, which could be associated with variations in treatment response and pharmacokinetics.

To perform the analysis, genes with unclassified annotations were excluded, resulting in 464,192 protein names out of 1,506,983 listed genes. Filtering out genes with a relative abundance below 1%, 1195 protein names remained. These gene/protein abundances were subjected to LEfSe analysis for each enterotype, and the genes/proteins associated with CRC/ADN cases with the highest effect size (LDA score >2) were identified. The number of differential genes (LDA >2) in CRC samples varied among enterotypes, ranging from 56 to 253, while for ADN samples, it ranged from 11 to 194.

In **Table 4.14**, the top 20 differentially expressed genes and microorganisms associated with each enterotype in the CRC group are listed. It was observed that genes from different organisms in different ETs are upregulated in cancerous conditions. In ET1, genes from Enterotoxigenic *Escherichia coli* species were predominantly identified, while genes from *Bacteroides* species dominated in ET2 and ET3.

Table 4.14: ET-specific CRC-associated genes and their species.

ET 1, CRC-associated genes	Species name
A7ZTI8_50S ribosomal protein L28*	<i>Escherichia coli</i> O139:H28
A8AMJ6_30S ribosomal protein S18*	<i>Citrobacter koseri</i>
A8AQG0_DNA_binding protein Fis*	<i>Citrobacter koseri</i>
C2ME56_Integrase core domain protein	<i>Porphyromonas uenonis</i>
D1PDG4_IstB_likeATP_binding protein	<i>Prevotella copri</i>
D1PH90_Transposase	<i>Prevotella copri</i>
D4JXT5_Flavodoxin_like fold	<i>Faecalibacterium prausnitzii</i>
P0A9P7_ATP_dependent RNA helicase DeaD*	<i>Escherichia coli</i> O6:H1
P0ADE3_Uncharacterized protein YtfK*	<i>Escherichia coli</i> O6:H1
P0AEP2_Galactose_proton symporter*	<i>Escherichia coli</i> O6:H1
P0C652_Insertion element IS1 protein InaA	<i>Escherichia coli</i>
P0CF06_Insertion element IS1 protein InaA*	<i>Escherichia coli</i>
P62524_ilv operon leader peptide*	<i>Escherichia coli</i> O157:H7
P68768_Cytosol amino peptidase*	<i>Escherichia coli</i> O157:H7
Q32IP8_IS1 ORF1*	<i>Shigella dysenteriae</i> serotype 1
ET 2, CRC-associated genes	Species name
B3JHP4_Group II intron_maturase_specific domain protein	<i>Bacteroides coprocola</i>
B3JNP7_Transposase	<i>Bacteroides coprocola</i>
B5A7F8_BUN24p15_5	<i>Bacteroides uniformis</i>
B5CUJ7_ISPg5_transposase Orf1	<i>Bacteroides plebeius</i>
B5CUJ8_Integrase core domain protein	<i>Bacteroides plebeius</i>
B5CXR8_ISPg5_transposase Orf2	<i>Bacteroides plebeius</i>
E4MCU7_Transposase_IS4 family	<i>Alistipes sp. HGB5</i>

Table 4.14: Continued.

ET 2, CRC-associated genes	Species name
G6AUA6_Transposase_IS116_IS110_IS902 family	<i>Prevotella stercorea</i>
G6AUA7_Transposase_IS116_IS110_IS902 family	<i>Prevotella stercorea</i>
I3YJF5_Site_specific recombinase XerD	<i>Alistipes finegoldii</i>
I3YJW5_Site_specific recombinase XerD	<i>Alistipes finegoldii</i>
I3YM16_50S ribosomal protein L3	<i>Alistipes finegoldii</i>
Q56VG1_TraH	<i>Bacteroides fragilis</i>
Q5LES4_Sulfate adenylyl transferase subunit 2	<i>Bacteroides fragilis</i>
ET 3, CRC-associated genes	Species name
A5ZWU2_Conserved domain protein	<i>Blautia obeum</i>
B5A7G0_Mob A	<i>Bacteroides uniformis</i>
B5A7G1_Mob B	<i>Bacteroides uniformis</i>
D7K6P6_RNA_directed DNA polymerase	<i>Bacteroides sp. 3_1_23</i>
E5WU41_Excisionase family DNA binding domain containing protein	<i>Bacteroides eggerthii</i> 1_2_48FAA
F0R9D1_Addiction module toxin_Txe_YoeB family	<i>Bacteroides salanitronis</i>
F3PHK0_Conserved domain protein	<i>Bacteroides clarus</i>
P30899_Beta_lactamase	<i>Bacteroides vulgatus</i>
P37247_Transposase for insertion sequence element IS4351	<i>Bacteroides fragilis</i>
Q5L8B9_50S ribosomal protein L14	<i>Bacteroides fragilis</i>
Q5L8Z8_Malate dehydrogenase	<i>Bacteroides fragilis</i>
Q5LAX0_Bifunctional protein FOLD	<i>Bacteroides fragilis</i>
Q5LI34_50S ribosomal protein L34	<i>Bacteroides fragilis</i>
Q89Z59_Probable type I restriction enzyme BthVORF4518P M protein	<i>Bacteroides thetaiotaomicron</i>
Q8A485_30S ribosomal protein S171	<i>Bacteroides thetaiotaomicron</i>
R5C195_Glycine_tRNA ligase	<i>Bacteroides sp.</i>
R5K610_Adenylo succinatelyase	<i>Bacteroides eggerthii</i>
W1I8Q3_Uncultured bacterium extra chromosomal DNARGF00115	<i>uncultured bacterium</i>

* Genes associated with both CRC and ADN.

Within ET1, the presence of a Flavodoxin-like protein belonging to *Faecalibacterium* species was noteworthy. This protein has been reported to play a critical role in the growth and survival of *Fusobacterium nucleatum*, a well-known CRC marker (Chen et al., 2020). Moreover, this gene has been considered a potential therapeutic target against gastric pathogens (Salillas & Sancho, 2020). In ET 2, *Bacteroides plebeius* was found to be increased from adenoma to carcinoma (p-value 0.026), also its transposase and integrase genes were found to be differentially increased in CRC. Similarly, beta-lactamase gene of *Bacteroides vulgatus* species was detected to be enriched in CRC.

The distinctiveness of the beta-lactamase enzyme, providing antibiotic resistance, in the context of CRC may be linked to the increase in antibiotic-resistant pathogens during dysbiosis. It has been reported that antibiotic-induced dysbiosis promotes cancer development and progression. Additionally, antibiotic resistance genes associated with CRC have been documented (Liu et al., 2021).

Table 4.15 presents the ET-specific genes and their associated organisms among the top 20 differentially expressed genes associated with ADN. Similar to **Table 4.14**, it is observed that genes from different organisms in different ETs are increased in adenoma conditions.

Table 4.15: ET-specific ADN-associated genes and their species.

ET 1, ADN-associated genes	Species name
A1A091_Translation initiation factor IF_1	<i>Escherichia coli O139:H28</i>
C0FZI4_CDP_alcohol phosphatidyl transferase	<i>Roseburia inulinivorans</i>
C4Z2R6_30S ribosomal protein S12	<i>Lachnospira eligens</i>
C4ZBS3_30S ribosomal protein S19	<i>Agathobacter rectalis</i>
C4ZFB9_Transcriptional regulator GltC	<i>Agathobacter rectalis</i>
D4JA34_ParB_like partition proteins	<i>Coprococcus catus</i>
D4JAB3_Glyoxalase_Bleomycin resistance protein_Dioxygenase superfamily	<i>Coprococcus catus</i>
D4JT21_DNA primase_Bacterial type_	<i>Eubacterium siraeum</i>
D4KKP9_Site_specific recombinase XerD	<i>Roseburia intestinalis</i>
D4KN84_Anti restriction protein	<i>Roseburia intestinalis</i>
D4L309_Response regulators consisting of a CheY_like receiver domain and a winged_helix DNA_binding domain	<i>Roseburia intestinalis</i>
D4LFS2_RNA polymerase sigma factor_sigma_70 family	<i>Ruminococcus sp.</i>
D4LKG3_Methylene_tetrahydrofolate reductase C terminal	<i>Ruminococcus sp.</i>
D4LLA5_Glyoxalase_Bleomycin resistance protein_Dioxygenase superfamily	<i>Ruminococcus sp.</i>
D4LMM9_50S ribosomal protein L3	<i>Ruminococcus sp.</i>
D4LR50_Monosaccharide ABC transporter ATP_binding protein CUT2 family_TC3_A_1_2	<i>Eubacterium obeum</i>
F3AZU8_Transposase	<i>Lachnospiraceae oral taxon 107</i>
Q8G7I6_Glucose_6_phosphate isomerase	<i>Bifidobacterium longum</i>
R6BEM2_L_aspartate oxidase	<i>Firmicutes bacterium</i>
ET 3, ADN-associated genes	Species name
D1PH90_Transposase*	<i>Prevotella copri</i>
D4KTC1_Transposase and inactivated derivatives	<i>Roseburia intestinalis</i>
D4KTC2_Transposase DDE domain	<i>Roseburia intestinalis</i>

Table 4.15: Continued.

ET 3, ADN-associated genes	Species name
D4LY53_Helix_turn_helix	<i>Eubacterium obeam</i>
D5HH09_H_Antiporter protein	<i>Coproccoccus sp.</i>
D6D1V0_DNA binding domain_excisionase family	<i>Bacteroides xylanisolvans</i>
E5VDI9_Glucosamine_fructose_6_phosphate aminotransferase	<i>Bacteroides sp.</i>
F0R0G6_IS66 Orf2 family protein	<i>Bacteroides salanitronis</i>
R6U0H8_Transcriptional regulator AraC family	<i>Bacteroides sp.</i>
ET 2, ADN-associated genes	Species name
A4WFH5_Site_specific DNA_methyl transferase adenine specific	<i>Enterobacter sp.</i>
A7ZT30_Universal stress protein B	<i>Escherichia coli O139:H28</i>
A7ZTI8_50S ribosomal protein L28	<i>Escherichia coli O139:H28</i>
A8AMJ6_30S ribosomal protein S18*	<i>Citrobacter koseri</i>
A8AQG0_DNA_binding protein Fis*	<i>Citrobacter koseri</i>
B5CW71_DNA binding domain_excisionase family	<i>Bacteroides plebeius</i>
D1PDZ5_Transposase_IS4 family	<i>Prevotella copri</i>
D6JB22_Predicted protein	<i>Escherichia coli B354</i>
P0A9P7_ATP_dependent RNA helicase DeaD*	<i>Escherichia coli O6:H1</i>
P0ADE3_Uncharacterized protein YtfK*	<i>Escherichia coli O6:H1</i>
P0AEP2_Galactose_proton symporter*	<i>Escherichia coli O6:H1</i>
P0CF06_Insertion element IS1 protein InaA*	<i>Escherichia coli</i>
P62524_ilv operon leader peptide*	<i>Escherichia coli O157:H7</i>
P68768_Cytosol amino peptidase*	<i>Escherichia coli O157:H7</i>
Q32IP8_IS1 ORF1*	<i>Shigella dysenteriae serotype 1</i>
Q8XBW9_Protein HokE	<i>Escherichia coli O157:H7</i>
W1AQL8_Mobile element protein	<i>absolute</i>
W1I7I5_Uncultured bacterium extra chromosomal DNARGI01876	<i>uncultured bacterium</i>

* Genes associated with both CRC and ADN

Among the top 20 genes with the highest LDA scores associated with CRC and ADN, Transposase proteins were found to be differentially increased in ADN and CRC samples in all three enterotypes. These proteins were found to be originating mostly from *Bacteroides* and *Prevotella* species, belonging to *Bacteroidales* order, which are the most abundant Gram-negative bacteria of healthy human gut microbiota (García-Bayona et al., 2021).

Transposons, which are mobile genetic elements, play a role in bacterial adaptation by inserting and removing themselves in genomic sequences. Addition and removal of these sequences is accomplished by transposase enzymes. Dysbiosis in the cancerous

state can facilitate increased horizontal gene transfer among gut microbiome members, promoting the spread of virulence factors and antibiotic resistance genes (Evanovich et al., 2022; Warnke-Sommer & Ali, 2017). In addition, overexpression of transposable elements has been associated with colorectal cancer progression (Zhu et al., 2021). The enriched transposase enzymes in ET1 and ET2 groups can be related with the dysbiotic effects of elevated ET-specific markers in the gut microbiota. *Lachnospiraceae bacterium* was notably enriched in adenoma (ADN) cases ($p=0.03$) when compared to controls, while *Bacteroides plebeius* abundance increased from adenoma to carcinoma sequence ($p=0.03$). Genes associated with both ADN and CRC have also been identified and are marked in both **Table 4.14** and **Table 4.15**. These shared differentially expressed genes, mainly found in *E. coli* strains suggesting that enterotoxigenic *E. coli* pathogens increase with the onset of early-stage cancer.

The ET-specific pathways associated with CRC and ADN were also identified, highlighting the distinct roles of different enterotypes. After filtering out 508 pathways with a relative abundance of less than 1%, a total of 208 pathways remained for further analysis. Using these 208 pathway abundances, the LEfSe analysis was performed for each enterotype. In the CRC-CTR comparison, 122, 13, and 61 pathways passed the $LDA>2$ filter for each ET, respectively. Similarly, in the ADN-CTR comparison, 35, 41, and 55 pathways passed the $LDA>2$ filter for each ET, respectively.

The top 10 differentially expressed pathways with the highest effect size (LDA score >2) associated with CRC and ADN cases are listed in **Table 4.16**. When comparing the ET-specific pathways among the top 20 pathways associated with CRC and ADN (**Figure 4.21**), it becomes evident that different mechanisms are involved in different ETs.

Table 4.16: The top 10 differentially expressed ET-specific pathways.

ET 1, CRC-associated pathways	Enriched in	Log(LDA)
ILEUSYN_PWY_L_isoleucine biosynthesis I_from threonine	CTR	2.87
VALSYN_PWY_L_valine biosynthesis	CTR	2.87
PWY_7111_pyruvate fermentation to isobutanol engineered	CTR	2.84
GLYCOGENSYNTH_PWY_glycogen biosynthesis I from ADP D Glucose	CTR	2.84
PWY_6151_S adenosyl L methionine cycle I	CTR	2.82

Table 4.16: Continued.

ET 1, CRC-associated pathways	Enriched in	Log(LDA)
PWY_6317_galactose degradation I_Leloir pathway	CTR	2.82
PWY66_422_D_galactose degradation V_Leloir pathway	CTR	2.82
PWY_621_sucrose degradation III_sucrose invertase	CTR	2.80
ARO_PWY_chorismate biosynthesis I	CTR	2.79
OANTIGEN_PWY_O_antigen building blocks biosynthesis_E_coli	CTR	2.78
ET 1, ADN-associated pathways	Enriched in	Log(LDA)
PWY_7219_adenosine ribonucleotides de novo biosynthesis	Adenoma	2.6846897 4
ANAGLYCOLYSIS_PWY_glycolysis III_from glucose	CTR	2.6687865 7
FERMENTATION_PWY_mixed acid fermentation	Adenoma	2.6519601 3
PWY_4981_L_proline biosynthesis II_from arginine	Adenoma	2.6356055 1
PWY4FS_8_phosphatidylglycerol biosynthesis II_non_plastidic	CTR	2.6291342 7
PWY4FS_7_phosphatidylglycerol biosynthesis I_plastidic	CTR	2.6291341 8
PANTO_PWY_phosphopantothenate biosynthesis I	Adenoma	2.5938926 9
PWY_5913_TCAcycleVI_obligate autotrophs	Adenoma	2.5761898
PWY0_1586_peptidoglycan maturation_meso_diaminopimelate containing	CTR	2.5685823 8
PWY_5104_L_isoleucine biosynthesis IV	Adenoma	2.5535855 4
ET 2, CRC-associated pathways	Enriched in	Log(LDA)
PWY_6151_S_adenosyl_L_methionine cycle I	CTR	2.42
PWY_6936_seleno_amino acid biosynthesis	CTR	2.35
PWY_6168_flavin biosynthesis III_fungi	CRC	2.33
PWY_1269_CMP_3_deoxy_D_manno_octulosonate biosynthesis I	CRC	2.30
PWY_6609_adenine andadenosine salvage III	CRC	2.26
PWY_6892_thiazole biosynthesis I_E_coli	CTR	2.22
GLUCONEO_PWY_gluconeogenesis I	CRC	2.20
TRPSYN_PWY_L_tryptophan biosynthesis	CRC	2.15
PWY_6507_4_deoxy_L_threo_hex_4_enopyranuronate degradation	CTR	2.09
PWY_5676_acetyl_CoA fermentation to butanoate II	CRC	2.09
ET 2, ADN-associated pathways	Enriched in	Log(LDA)
PWY_6168_flavin biosynthesis III_fungi	CTR	2.97

Table 4.16: Continued.

ET 2, ADN-associated pathways	Enriched in	Log(LDA)
PWY0_162_superpathway of pyrimidine ribonucleotides de novo biosynthesis	CTR	2.78
THISYNARA_PWY_superpathway of thiamin diphosphate biosynthesis III eukaryotes	Adenoma	2.76
P124_PWY_Bifidobacterium hunt	Adenoma	2.76
CENTFERM_PWY_pyruvate fermentation to butanoate	CTR	2.74
PWY4FS_7_phosphatidylglycerol biosynthesis I plastidic	CTR	2.74
COA_PWY_coenzyme A biosynthesis I	Adenoma	2.73
PWY0_1319_CDP_diacylglycerol biosynthesis II	CTR	2.66
PWY_5189_tetrapyrrole biosynthesis II_from glycine	CTR	2.66
ET 3, CRC-associated pathways	Enriched in	Log(LDA)
PWY_6737_starch degradation V	CTR	2.94
PWY_6122_5_amino imidazole ribonucleotide biosynthesis II	CTR	2.88
PWY_6277_superpathway of 5_amino imidazole ribonucleotide biosynthesis	CTR	2.88
PWY_7228_superpathway of guanosine nucleotides de novo biosynthesis I	CRC	2.88
TRNA_CHARGING_PWY_tRNA charging	CTR	2.87
PWY_6125_superpathway of guanosine nucleotides de novo biosynthesis II	CRC	2.81
PWY_6121_5_amino imidazole ribonucleotide biosynthesis I	CTR	2.80
COMPLETE_ARO_PWY_superpathway of aromatic amino acid biosynthesis	CTR	2.76
PWY_7111_pyruvate fermentation to isobutanol engineered	CTR	2.76
ILEUSYN_PWY_L_isoleucine biosynthesis I_from threonine	CTR	2.74
ET 3, ADN-associated pathways	Enriched in	Log(LDA)
ANAGLYCOLYSIS_PWY_glycolysis III_from glucose	CTR	2.89
PWY_3841_folate transformations II	CTR	2.87
PWY_1042_glycolysis IV_plant cytosol	CTR	2.87
COA_PWY_1_coenzyme A biosynthesis II_mammalian	CTR	2.85
PWY_7664_oleate biosynthesis IV_anaerobic	Adenoma	2.85
f_1CMET2_PWY_N10_formyl_tetrahydrofolate biosynthesis	CTR	2.81
FASYN_ELONG_PWY_fattyacid elongation_saturated	Adenoma	2.77
PWY0_862_5Z_dodec_5_enoate biosynthesis	Adenoma	2.74
PHOSLIPSYN_PWY_superpathway of phospholipid biosynthesis I bacteria	Adenoma	2.73
PWY_6385_peptidoglycan biosynthesis III_mycobacteria	CTR	2.69

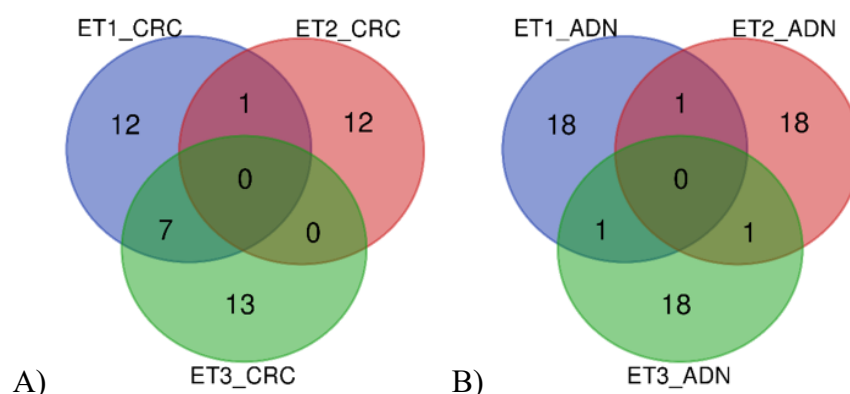


Figure 4.21: Comparison of the top 20 most effective (LDA>2) pathways in ETs associated with A) CRC and B) ADN.

4.8. Correlation Analysis of ET-Specific Microbial Markers Associated with BMI and Age

Correlation analysis was conducted to examine the relationship between microbial markers and age and obesity groups. A statistically significant negative rho value ($p<0.05$) indicates an inverse relationship among variables, while a positive value indicates a direct linear relationship. A rho value between 0.6 and 1 indicates a high correlation. Based on this information, the microbial markers identified in this study were examined for their association with age and obesity groups, and markers with high correlation are listed in **Table 4.17**.

The correlation analysis focused on determining the microbial markers that were statistically significantly associated with age and obesity groups. Correlation coefficients were measured separately for CRC, CTR, and ADN samples, and all significant coefficients ($p<0.05$) were analyzed. The identified microbial markers and their correlation with age and BMI (body mass index) were documented. Age groups were categorized into different ranges (40<, 41-50, 51-60, 61-70, 71-80, and 81-90), while BMI groups included <18.5 (thin), 18.5-24.9 (healthy), 25-29.9 (overweight), and >30 (obese) groups.

Table 4.17: Results with high correlation between age, BMI and microbial markers.

	Microbial Marker Name	ADN (rho)	CRC (rho)	CTR (rho)
ET 1	Age (<=40) (n=16) Correlation			
	<i>s__Clostridium_symbiosum</i>	NA	-0,921	0,218
	<i>s__Parvimonas_micra</i>	NA	-0,918	NA
	<i>s__Streptococcus_thermophilus</i>	NA	-0,918	-0,204
	<i>s__Solobacterium_moorei</i>	NA	-0,918	-0,134
	<i>s__Peptostreptococcus_stomatis</i>	NA	-0,918	NA
	Age (71-80) (n=100) Correlation			
	<i>s__Ruminococcus_sp_5_1_39BFAA</i>	0,654	NS	-0,292
	Lean (BMI < 18,5) (n=14) Correlation			
	<i>s__Peptostreptococcus_stomatis</i>	NA	0,683	NA
	<i>s__Adlercreutzia_equolifaciens</i>	NA	-0,635	NA
ET 2	Age (81-90) (n=14) Correlation			
	<i>s__Eubacterium_ventriosum</i>	NA	0,608	NA
	Obese (BMI> 30) (n=30) Correlation			
	<i>s__Phascolarctobacterium_succinatutens</i>	NS	NS	-0,813
ET 3	Age (<=40) (n=19) Correlation			
	<i>s__Ruminococcus_gnavus</i>	NA	NA	0,613
	Age (41-50) (n=18) Correlation			
	<i>s__Fusobacterium_mortiferum</i>	NA	0,801	NA
	<i>s__Parvimonas_micra</i>	NA	-0,741	NS
	<i>s__Gemella_morbillorum</i>	NA	-0,741	NA
	<i>s__Porphyromonas_uenonis</i>	NA	-0,741	NA
	<i>s__Bacteroidales_bacterium_ph8</i>	NA	NS	-0,945
	Age (51-60) (n=47) Correlation			
	<i>s__Ruminococcus_torques</i>	-0,946	NS	NS

NA: Correlation was not calculated for those with n≤3 sample numbers.

NS: No statistically significant (p<0.05) correlation was observed.

The correlation analysis revealed that, for each enterotype, there were no significant high correlations between age and microbial markers. However, high correlations were observed within different age groups. Accordingly, in the 1st enterotype, there was a high negative correlation observed between *Clostridium symbiosum*, *Parvimonas micra*, *Solobacterium moorei*, *Peptostreptococcus stomatis*, *Streptococcus thermophilus* species and the age category below 40. For the same enterotype, a positive relationship was observed between the relative abundance of *Ruminococcus sp. 5_1_39BFAA* and Adenoma patients aged 71-80. These findings provide valuable insights into the identification of microbial markers that may be more relevant for specific age groups. Through the correlation analysis conducted with BMI, a

significant interaction was observed between underweight (BMI<18.5) cancer patients and *Peptostreptococcus stomatis* and *Adlercreutzia equolifaciens* species.

4.9. Microbial Gene/Protein Targeted Drug Screening

No targeted drug molecules were identified for CRC and ADN-related genes (**Table 4.14** and **Table 4.15**) in the DrugBank database. However, when homologs of these proteins were searched, matching targeted drug molecules were found (listed in the **Appendix A** and **Appendix B**). Some of these drugs are currently in clinical trials for treating different diseases. These drug molecules can be further investigated for potential use in microbial gene/protein discovery and drug repositioning for colon cancer treatment. The findings represent preliminary research for advanced microbial gene therapies.

5. DISCUSSION

This is the first enterotype-based meta-analysis of fecal metagenomic samples specific for CRC and ADN samples at species resolution. There have been few studies focusing on the changes in the gut microbiome of CRC or ADN patients stratified into different enterotypes (Cai et al., 2023; Lv et al., 2023; Yang et al., 2019.), however no distinct microbial species specific for enterotypes were proposed as diagnostic markers.

Our study employed enterotype-specific analysis to uncover distinctive microbial markers associated with different enterotypes. Even with the stratification of metagenomes into enterotypes, the detection of well-known colorectal cancer (CRC) markers, including *Fusobacterium nucleatum* and *Parvimonas stomatis* (Wu et al., 2021), along with the common marker *Gemella morbillorum* crucial for adenoma-carcinoma differentiation, (Feng et al., 2015) highlights their robustness for microbiome-based CRC diagnosis development. This underscores the potential universality of these markers, irrespective of enterotype differentiation.

Our analysis not only identified common markers but also revealed enterotype-specific markers at species resolution with significant differences. The identification of *Ruminococcus* and *Eubacterium* genera, recognized as butyrate-producing bacteria along with other taxa like *Firmicutes*, *Clostridium*, and *Coprococcus* (Takahashi et al., 2016; Zhu et al., 2021) as specific markers for Enterotype 1 (ET1) is a key finding. Butyrate, derived from the microbiota, plays a crucial role in enhancing the intestinal mucosal immune barrier, regulating systemic immune responses, and preventing infections (Zhu et al., 2021). Consequently, an imbalance in butyrate production may lead to dysbiosis. Key species which contribute to butyrate production in the gut such as *Eubacterium rectale*, *Ruminococcus callidus*, *Ruminococcus sp 5 1 39BFAA*, *Coprococcus comes*, *Bifidobacterium* (*B. bifidum*, *B. catenulatum*, *B. pseudocatenulatum*, *B. longum*), and *Gordonibacter pamelaee* were identified as CRC-depleted markers in our study. These bacteria play a vital role in fermenting dietary fibers and resistant starches, ultimately producing butyrate—an essential energy source for colonocytes with various beneficial effects on intestinal health. Since it is known that loss of beneficial microbial organisms is one of the reasons of gut dysbiosis, a notable decrease in differential abundance of these species in people with cancer makes them to serve as selective markers for CRC patients stratified into ET1,

and they could also be considered as dietary supplements in CRC management. Furthermore, beneficial genera such as *Ruminococcus*, *Bifidobacterium*, *Lachnospira*, *Clostridiales*, and *Lactobacillus*, known for their anti-inflammatory properties, have been patented as pharmaceutical compositions for the treatment of dysbiosis (Berry, et al., 2015a; Berry, et al., 2015b). Administration of *B. longum* as an oral supplement has been shown to suppress colon tumor incidence and tumor volume (Singh et al., 1997).

Given that dietary factors play a crucial role in the majority of CRC cases and impact cancer progression (Pericleous et al., 2013), probiotics are being explored as potential adjuncts in treatment (Torres-Maravilla et al., 2021). There have been recent studies indicating an association with dietary patterns and enterotypes (Wu et al., 2011a). People with dominated *Ruminococcus* species in their gut microbiome have been associated with having a diet high in resistant starch, where these species play an important role in releasing energy through the degradation of starch (Arumugam et al., 2011; Ze et al., 2012).

Notably, *Gordonibacter pamelaee* was depleted in CRC samples from the ET1 group compared to both healthy and adenoma samples. This anti-inflammatory bacterium has been linked to Inflammatory Bowel Diseases, contributing to the modification of bile acids in the gut microbiome (Lee et al., 2021). In this study, it emerged as a novel *Ruminococcus-Eubacterium* enterotype-specific marker for colon cancer.

Desulfovibrio desulfuricans is another novel microbial marker for *Ruminococcus* dominated enterotype. While the *Desulfovibrio* genus was observed to be abundant in colon cancer samples in various studies (Xu et al., 2020), *D. desulfuricans*, a specific species within this genus, was not previously identified as enriched in CRC cases. However, recent reports have indicated its enrichment in patients with Inflammatory Bowel Disease (IBD), a population with an increased likelihood of developing colorectal cancer (Khaledi et al., 2024).

For ET2, *Coprobacter fastidiosus*, *Bacteroides intestinalis*, *Megamonas hypermegale* and *Burkholderiales bacterium 1_1_47* were found to be CRC markers depleted in cancer cases of people stratified into *Bacteroidetes* dominated gut microbiota. *Anaerotruncus colihominis*, *Eubacterium biforme*, *Clostridium scindens*,

Granulicatella adiacens, *Veillonella atypica*, *Holdemania filiformis*, *Coprococcus catus* were enriched in CRC when compared to healthy controls.

It has been shown that the fecal microbial community in humans who consume protein and animal fat for a long period is dominated by the *Bacteroides* genus (Choi et al., 2021). Bile acids (BAs) are synthesized as a result of cholesterol catabolism in the liver tissue and involved in the absorption of fat and the regulation of lipid homeostasis (Qi et al., 2015). The transformation of bile acids into secondary bile acids is crucially influenced by the gut microbiota. These secondary bile acids are involved in the dysbiosis of the gut microbiome and have a role in promoting CRC (Liu et al., 2022).

Clostridium scindens is recognized as a prominent bacterium in the production of secondary bile acids through the 7 α / β -dehydroxylation pathway (Chattopadhyay et al., 2022; Ridlon & Hylemon, 2012). Several studies have suggested its potential as a marker for IBD and CRC cases (Wang et al., 2023; Wu et al., 2021). To the best of our knowledge, our study is the first to propose this species as a specific CRC marker within the *Bacteroides* enterotype in the literature.

Holdemania filiformis has been implicated in bile acid transformation and the maintenance of gut homeostasis (Su et al., 2022). Additionally, it has been identified at elevated levels in individuals exhibiting positive responses to immunotherapies, suggesting an immunomodulatory impact. Recent studies have highlighted the gut microbiota's impact on clinical responses to immunotherapy. *F. prausnitzii* and *H. filiformis* were found in higher abundance in the gut microbiota of melanoma patients who responded to Ipilimumab (anti-CTLA-4) + Nivolumab (anti-PD-1) immunotherapies (Aarnoutse et al., 2019; Oh et al., 2021). Since these immune checkpoint inhibitors are also utilized in colorectal cancer treatment (Jácome & Eng, 2019), this emphasizes *H. filiformis*'s potential as a specific biomarker for precise patient selection in cancer immunotherapy. However, it is essential to acknowledge that further research is needed to illuminate and substantiate this hypothesis.

While a limited number of studies have noted the distinct abundance of species like *H. filiformis*, *G. adiacens*, *V. atypica*, *A. coli*hominis, *C. catus* in cancer cases (Alexander et al., 2023; Thomas et al., 2019; Uppakarn et al., 2021) and depletion of *C. fastidiosus* and *Burkholderiales bacterium I_1_47* (Avuthu & Guda, 2022; Coker et al., 2022),

our study marks the first proposal of these markers as *Bacteroides* enterotype-specific CRC markers.

A plant-rich diet, abundant in carbohydrates and fibres, has been linked to the *Prevotella*-dominated enterotype (Wu et al., 2011b), dominated by fibre-using bacteria that ferment dietary fibre into short chain fatty acids (SCFAs) such as acetate, butyrate, and propionate. These SCFAs, especially propionate, play a key role in regulating metabolic processes in *Prevotella* enterotype. Bacterial species from both *Bacteroides* and *Prevotella* genera produce propionate, while especially *Prevotella* seems to effectively produce propionate in larger amounts (Christensen et al., 2018). Administration of SCFAs has been shown to inhibit the expression of pro-inflammatory cytokines, thereby preventing tumour development and reducing colon inflammation (Tian et al., 2018). *Alistipes* and *Bacteroides* species are known to be among SCFA (propionate and acetate) producing gut bacteria (Zhang et al., 2023). In our study, three *Alistipes* species (*A. putredinis*, *A. senegalensis*, *Bacteroidales bacterium ph8* (*Alistipes obesi* ph8)) and *Bacteroides coprocola* were found to be depleted in CRC samples clustered in *Prevotella* enterotype.

Alistipes is a relatively new genus isolated from human feces recently (Shkoporov et al., 2015). Contrary to earlier findings that identified *Alistipes* species as enriched in CRC without considering enterotype (Feng et al., 2015), our study reveals propionate producing *Alistipes* species reduction in CRC cases. *Alistipes* has demonstrated both protective and detrimental effects on various diseases. While its pathogenic impact on CRC has been noted, there are indications of beneficial immunomodulatory effects in cancer immunotherapy (Parker et al., 2020). Importantly, *Alistipes* reduction was found to be associated with parallel reduction in positive responses to cancer immunotherapy (Iida et al., 2013). Our study sheds light on the complex role of *Alistipes* in CRC and its potential implications for cancer immunotherapy.

For *Prevotella* enterotype, RF classification identified *Streptococcus mitis oralis pneumoniae*, *Veillonella parvula* and *Dorea formicigenerans* species as ET specific CRC enriched markers within the top importance (**Figure 4.6**). *Veillonella* genus has been associated with CRC in *Prevotella*-dominated enterotypes (T. Yang et al., n.d.) although no detailed species-level representation was conducted. Our analysis identified *V. parvula* as a novel and significant CRC marker.

D. formicigenerans has been identified as a differential species in CRC in some studies (Shuwen et al., 2022; X. Wu et al., 2024), lacking a specific focus. Our research further emphasizes this species as a potent CRC marker specific to the *Prevotella* enterotype. *D. formicigenerans* was also found to be prevalent in the gut microbiota of individuals responding to cancer immunotherapies utilized in CRC treatment (Aarnoutse et al., 2019). This suggests the potential of *D. formicigenerans* as a specific microbial marker for patient selection in immunotherapy, particularly within the *Prevotella* enterotype. It's important to note that further supporting research is required to validate these findings.

Microbial differences along the adenoma-carcinoma sequence were also investigated to determine markers predicting cancer progression. Specific markers for precancerous adenoma cases within Enterotypes were pinpointed, suggesting potential applications in early diagnosis and preventive therapy. Notably, *Actinomyces* sp. ICM47 and *Enterobacter aerogenes* exhibited a significant increase from healthy to adenoma in ET1 and ET2, respectively (**Table 4.8**). Additionally, unique ET1 markers *Odoribacter splanchnicus*, *Parabacteroides merdae*, *Alistipes finegoldii*, and *Escherichia coli* were identified as highly important in distinguishing adenoma from carcinoma (**Figure 4.13**). Within the *Bacteroides* enterotype (ET2), *Streptococcus sanguinis*, *Eubacterium ramulus*, and *Ruminococcus obeum* were recognized as colorectal cancer (CRC) markers compared to adenoma, prominently featured in the RF classifier (**Figure 4.13**). Within the ET3 group, where the *Prevotella* genus is predominant, *Streptococcus thermophilus* and *Ruminococcus gnavus* emerged as specific markers enriched in CRC compared to adenoma, once again demonstrating high significance (**Figure 4.13**). Among these markers, *E. coli*, *E. aerogenes* were detected previously as adenoma related species that can convert cysteine to Hydrogen sulfide (H₂S), ammonia, and pyruvate through cysteine desulfhydrase. These compounds are known to play a role in modulating inflammation and contribute to toxic activities that promote carcinogenesis (Vacante et al., 2020). Also, recent studies indicated that *E. coli* and *Enterobacter* strains might promote colon cancer initiation and progression through production of toxins (El Tekle & Garrett, 2023). *O. splanchnicus* was also identified as CRC driver bacteria (P. Cai et al., 2023). *Alistipes* species were reported to play a critical role in inflammation-related carcinogenesis, inducing tumorigenesis (Tilg et al., 2018) and *A. finegoldii* has been increasingly

detected in a mouse CRC model following the "adenoma-carcinoma sequence (Sun et al., n.d.). *R. gnavus* was depleted in both cancer and adenoma cases when compared to controls in *Prevotella* enterotype.

These markers specific to enterotypes hold potential for colon cancer diagnosis and contribute significantly to the pathogenesis of CRC. Moreover, there is an opportunity to utilize them in preventive therapies, aiming to modulate the gut microbiota, restore balance, and mitigate microbial dysbiosis.

Given that profiling microbial communities using 16S rRNA genes is a simpler and more cost-efficient method for identifying the taxonomic composition of a microbial community, it is crucial to assess whether the 16S approach can serve as an effective alternative to WGS approach for non-invasive colon cancer screening. A comparison between these two approaches was conducted to determine if there is any consistency between them. There were not much consistency between WGS and 16S datasets since they include different sample size, sequencing depth in diversity of populations and they inherit differences in the both sequencing method and analytical approaches. 16S rRNA sequencing typically provides lower taxonomic resolution compared to WGS, which provide higher resolution down to the species or even strain level. Thus this comprehensive approach allows more precise taxonomic identification. 16S rRNA sequencing infers relative abundances of taxa based on 16S rRNA gene amplification, which can be biased by differences in gene copy number and PCR efficiency, whereas shotgun metagenomics (WGS) provides more accurate quantitative data on microbial abundances by sequencing the entire genetic material without relying on PCR amplification. WGS also enables characterization of functional genes, pathways, and the overall functional potential of the microbiome, offering insights into metabolic activities and potential interactions with the host.

In 16S data analysis, some of the sequencing artifacts can be eliminated during the bioinformatics steps. For this purpose, the amplicon sequence variant (ASV) identification (Callahan et al., 2016) technique was used in this thesis, which identifies unique sequences by removing low-quality reads and sequence artifacts through a probabilistic model that assesses the likelihood of a rare sequence being a true biological variant. This technique is superior to operational taxonomic unit (OTU) clustering, where similar sequences are grouped together based on percent sequence similarity.

Even though there are above mentioned limitations to determine consistent microbial markers using WGS and 16S datasets, we were able to identify some ET-specific microbial markers consistently detected using the two approaches. *Ruminococcus callidus*, which is one of the key species which contribute to butyrate production, was depleted in CRC cases with *Eubacterium-Ruminococcus* dominated enterotype, as a result of the meta-analysis of both 16S and WGS datasets. *Bacteroides salyersiae* was also commonly detected to be enriched in CRC cases in the same ET group. *B. salyersiae* is normally found in healthy, however it can be an opportunistic pathogen under certain conditions and can cause infections when the body's immune system is compromised, which may explain why it was detected as enriched state in CRC samples.

Several studies have compared 16S and WGS results to detect consistency. Zeller et al., 2014 compared CRC marker species obtained from fecal shotgun sequencing and fecal 16S sequencing. They performed shotgun sequencing on 141 individuals and 16S sequencing on 116 individuals from the same cohort. Despite sampling from the same individuals, the consistency between the two methods was not very high. However, there were consistent results in identifying *Fusobacterium nucleatum*, *Clostridium symbiosum*, and *Bacteroides fragilis* as enriched in CRC, while *Eubacterium rectale*, *E. ventriosum*, *E. eligens*, *Bacteroides caccae*, and *Streptococcus salivarius* were depleted in CRC in both 16S and WGS analyses. They also trained separate classifiers to determine microbial markers for CRC detection and found similar accuracies (AUC 0.84 using WGS, AUC 0.82 using 16S). However, most of the distinguishing taxa in the 16S results were not at the species level resolution.

In our results, WGS-based Random Forest classifiers for CRC and ADN detection achieved generally better accuracies when compared to the 16S-based classifiers, except that ET2 classifiers had similar accuracies with 0.73 and 0.72 AUC scores for WGS-based and 16S-based analysis, respectively. To establish the viability and cost-effectiveness of 16S rRNA sequencing as an alternative to WGS for non-invasive colon cancer screening, and to identify robust microbial markers detectable by both methods, future studies should involve larger, more diverse cohorts, and validate findings using external datasets.

There are also studies (Feng et al., 2015; Wirbel et al., 2019a) highlighted the use of shotgun metagenomic sequencing to uncover species-level differences and functional

potential in the microbiome of CRC patients, which were not able to be captured by 16S rRNA sequencing. Recent studies have focused on developing strategies to map detected taxa in shotgun data to the nearest corresponding 16S taxa (Ramon et al., 2024). However, substituting the shotgun model's microbial signature with the 16S taxa selected through this mapping results in decreased model accuracy. Similar research has also involved training separate models with shotgun and 16S data, with shotgun data consistently showing better model performance (Hannigan et al., 2018; Nagata et al., 2022; Vogtmann et al., 2016). As a result, since each technology has its own advantages and disadvantages (Peterson et al., 2021), integrating both approaches can offer a more comprehensive understanding of the microbiome's role in diseases such as colorectal cancer.

REFERENCES

- Aarnoutse, R., Ziemons, J., Penders, J., Rensen, S. S., De Vos-Geelen, J., & Smidt, M. L. (2019). The Clinical Link between Human Intestinal Microbiota and Systemic Cancer Therapy. *International Journal of Molecular Sciences*, 20(17).
- Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S. T., Methé, B., Schloss, P. D., Gevers, D., Mitreva, M., & Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computational Biology*, 8(6).
- Alexander, J. L., Posma, J. M., Scott, A., Poynter, L., Mason, S. E., Doria, M. L., Herendi, L., Roberts, L., McDonald, J. A. K., Cameron, S., Hughes, D. J., Liska, V., Susova, S., Soucek, P., der Sluis, V. H. van, Gomez-Romero, M., Lewis, M. R., Hoyles, L., Woolston, A., ... Kinross, J. (2023). Pathobionts in the tumour microbiota predict survival following resection for colorectal cancer. *Microbiome*, 11(1).
- Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. FastQC: A Quality Control Tool for High Throughput Sequence Data. <https://www.scienceopen.com/document?vid=de674375-ab83-4595-afa9-4c8aa9e4e736>.
- Arumugam, M., Raes, J., Pelletier, E., Paslier, D. Le, Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., ... Zeller, G. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346), 174.
- Avuthu, N., & Guda, C. (2022). Meta-Analysis of Altered Gut Microbiota Reveals Microbial and Metabolic Biomarkers for Colorectal Cancer. *Microbiology Spectrum*, 10(4).
- Baxter, N. T., Ruffin, M. T., Rogers, M. A. M., & Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, 8(1), 1–10.
- Berry, D., Afeyan, N. B., Kaplan, J., & Rahman, S. (2015a). “Probiotic and prebiotic compositions, and methods of use thereof for modulation of the microbiome”, U.S. Patent 10980845B2.
- Berry, D., Kaplan, J., & Rahman, S. (2015b). “Probiotic compositions containing clostridiales for inhibiting inflammation”, U.S. Patent 9610307.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E.,

- Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. In *Nature Biotechnology* (Vol. 37, Issue 8, pp. 852–857). Nature Publishing Group.
- Brenner, H., Kloor, M., & Pox, C. P. (2014). Colorectal cancer. *Lancet (London, England)*, 383(9927), 1490–1502.
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60.
- Bullman, S., Pedomallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., Neuber, D., Huang, K., Guevara, F., Nelson, T., Chipashvili, O., Hagan, T., Walker, M., Ramachandran, A., Diosdado, B., Serna, G., Mulet, N., Landolfi, S., Ramon, S., ... Meyerson, M. (2017). Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science*, 358(6369), 1443–1448.
- Bushman, F. D., Lewis, J. D., & Wu, G. D. (2013). Diet, gut enterotypes and health: is there a link? *Nestle Nutrition Institute Workshop Series*, 77, 65–73.
- Cai, J. A., Zhang, Y. Z., Yu, E. Da, Ding, W. Q., Jiang, Q. W., Cai, Q. C., & Zhong, L. (2023). Gut Microbiota Enterotypes Mediate the Effects of Dietary Patterns on Colorectal Neoplasm Risk in a Chinese Population. *Nutrients*, 15(13).
- Cai, P., Xiong, J., Sha, H., Dai, X., & Lu, J. (2023). Tumor bacterial markers diagnose the initiation and four stages of colorectal cancer. *Frontiers in Cellular and Infection Microbiology*, 13, 1123544.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583.
- Chattopadhyay, I., Gundamaraju, R., Jha, N. K., Gupta, P. K., Dey, A., Mandal, C. C., & Ford, B. M. (2022). Interplay between Dysbiosis of Gut Microbiome, Lipid Metabolism, and Tumorigenesis: Can Gut Dysbiosis Stand as a Prognostic Marker in Cancer? *Disease Markers*, 2022.
- Chen, S., Su, T., Zhang, Y., Lee, A., He, J., Ge, Q., Wang, L., Si, J., Zhuo, W., & Wang, L. (2020). *Fusobacterium nucleatum* promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7. *Gut Microbes*, 11(3), 511–525.
- Choi, S. I., Kim, N., Nam, R. H., Park, J. H., Nho, H., Yu, J. E., Song, C.-H., Lee, S. M., & Lee, D. H. (2021). Fecal Microbial Enterotypes Differentially Respond to a High-fat Diet Based on Sex in Fischer-344 Rats. *Journal of Cancer Prevention*, 26(4), 277.
- Christensen, L., Roager, H. M., Astrup, A., & Hjorth, M. F. (2018). Microbial enterotypes in personalized nutrition and obesity management. *The American Journal of Clinical Nutrition*, 108(4), 645–651.

- Chung, L., Orberg, E. T., Geis, A. L., Wu, S., Sears, C. L., & Housseau, F. (2018). *Bacteroides fragilis* toxin coordinates a pro-carcinogenic inflammatory cascade via targeting of colonic epithelial cells. *Cell.ComL Chung, ET Orberg, AL Geis, JL Chan, K Fu, CEDS Shields, CM Dejea, P Fathi, J ChenCell Host & Microbe*, 2018•*cell.Com*, 23, 203–214.
- Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O’connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O’sullivan, O., Fitzgerald, G. F., Deane, J., O’connor, M., Harnedy, N., O’connor, K., O’mahony, D., Van Sinderen, D., Wallace, M., Brennan, L., ... O’toole, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* 2012 488:7410, 488(7410), 178–184.
- Coker, O. O., Liu, C., Wu, W. K. K., Wong, S. H., Jia, W., Sung, J. J. Y., & Yu, J. (2022). Altered gut metabolites and microbiota interactions are implicated in colorectal carcinogenesis and can be non-invasive diagnostic biomarkers. *Microbiome*, 10(1).
- Costea, P. I., Hildebrand, F., Manimozhiyan, A., Bäckhed, F., Blaser, M. J., Bushman, F. D., De Vos, W. M., Ehrlich, S. D., Fraser, C. M., Hattori, M., Huttenhower, C., Jeffery, I. B., Knights, D., Lewis, J. D., Ley, R. E., Ochman, H., O’Toole, P. W., Quince, C., Relman, D. A., ... Bork, P. (2018). Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1), 8–16.
- Cougnoux, A., Delmas, J., Gibold, L., Faïs, T., Romagnoli, C., Robin, F., Cuevas-Ramos, G., Oswald, E., Darfeuille-Michaud, A., Prati, F., Dalmasso, G., & Bonnet, R. (2016). Small-molecule inhibitors prevent the genotoxic and protumoural effects induced by colibactin-producing bacteria. *Gut*, 65(2), 278–285.
- Csárdi, G., & Nepusz, T. (2006). *The igraph software package for complex network research*.
- Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K. K., Zhao, L., Chen, Z., Chan, F. K. L., Kristiansen, K., Sung, J. J. Y., Wong, S. H., & Yu, J. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome*, 6(1), 70.
- Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., & Wallace, M. B. (2019). Colorectal cancer. *Lancet (London, England)*, 394(10207), 1467–1480.
- Dethlefsen, L., & Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(SUPPL. 1), 4554–4561.
- D’Haens, G. R., & Jobin, C. (2019). Fecal Microbial Transplantation for Diseases Beyond Recurrent *Clostridium Difficile* Infection. In *Gastroenterology* (Vol. 157, Issue 3, pp. 624–636).

- Dikeocha, I. J., Al-Kabsi, A. M., Chiu, H. T., & Alshawsh, M. A. (2022). Faecalibacterium prausnitzii Ameliorates Colorectal Tumorigenesis and Suppresses Proliferation of HCT116 Colorectal Cancer Cells. *Biomedicines*, 10(5).
- dos Reis, S. A., da Conceição, L. L., Siqueira, N. P., Rosa, D. D., da Silva, L. L., & Peluzio, M. do C. G. (2017). Review of the mechanisms of probiotic actions in the prevention of colorectal cancer. In *Nutrition Research* (Vol. 37, pp. 1–19). Elsevier Inc.
- El Tekle, G., & Garrett, W. S. (2023). Bacteria in cancer initiation, promotion and progression. *Nature Reviews. Cancer*, 23(9), 600–618.
- Evanovich, E., Mattos, P. J. de S. M., & Guerreiro, J. F. (2022). Comprehensive Comparative Genomic reveals: Bacteroides fragilis is a reservoir of antibiotic resistance genes in the gut microbiota. *BioRxiv*, 2022.05.30.494044.
- Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J. Y., ... Wang, J. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature Communications*, 6, 6528–6528.
- Flemer, B., Lynch, D. B., Brown, J. M. R., Jeffery, I. B., Ryan, F. J., Claesson, M. J., O’Riordain, M., Shanahan, F., & O’Toole, P. W. (2017). Original article: Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*, 66(4), 633.
- Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11), 962–968.
- García-Bayona, L., Coyne, M. J., & Comstock, L. E. (2021). Mobile Type VI secretion system loci of the gut Bacteroidales display extensive intra-ecosystem transfer, multi-species spread and geographical clustering. *PLoS Genetics*, 17(4).
- Gupta, A., Dhakan, D. B., Maji, A., Saxena, R., P.K., V. P., Mahajan, S., Pulikkan, J., Kurian, J., Gomez, A. M., Scaria, J., Amato, K. R., Sharma, A. K., & Sharma, V. K. (2019). Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *MSystems*, 4(6).
- Hale, V. L., Chen, J., Johnson, S., Harrington, S. C., Yab, T. C., Smyrk, T. C., Nelson, H., Boardman, L. A., Druliner, B. R., Levin, T. R., Rex, D. K., Ahnen, D. J., Lance, P., Ahlquist, D. A., & Chia, N. (2017). Shifts in the fecal microbiota associated with adenomatous polyps. *Cancer Epidemiology Biomarkers and Prevention*, 26(1), 85–94.
- Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., Koumpouras, C. C., & Schloss, P. D. (2018). Diagnostic potential and interactive dynamics of the colorectal cancer virome. *MBio*, 9(6).

- Hewitson, P., Glasziou, P., Irwig, L., Towler, B., & Watson, E. (2007). Screening for colorectal cancer using the faecal occult blood test, Hemoccult. *Cochrane Database of Systematic Reviews*, 1.
- Iida, N., Dzutsev, A., Stewart, C. A., Smith, L., Bouladoux, N., Weingarten, R. A., Molina, D. A., Salcedo, R., Back, T., Cramer, S., Dai, R. M., Kiu, H., Cardone, M., Naik, S., Patri, A. K., Wang, E., Marincola, F. M., Frank, K. M., Belkaid, Y., ... Goldszmid, R. S. (2013). Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science*, 342(6161), 967–970.
- Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Levin, T. R., Lavin, P., Lidgard, G. P., Ahlquist, D. A., & Berger, B. M. (2014). Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *New England Journal of Medicine*, 370(14), 1287–1297.
- Jácome, A. A., & Eng, C. (2019). Role of immune checkpoint inhibitors in the treatment of colorectal cancer: focus on nivolumab. *Expert Opinion on Biological Therapy*, 19(12), 1247–1263.
- Jiang, P., Wu, S., Luo, Q., Zhao, X., & Chen, W.-H. (2021). Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models. *MSystems*, 6(3).
- Kaźmierczak-Siedlecka, K., Daca, A., Fic, M., van de Wetering, T., Folwarski, M., & Makarewicz, W. (2020). Therapeutic methods of gut microbiota modification in colorectal cancer management – fecal microbiota transplantation, prebiotics, probiotics, and synbiotics. *Gut Microbes*, 11(6), 1518.
- Khaledi, M., Poureslamfar, B., Alsaab, H. O., Tafaghodi, S., Hjazzi, A., Singh, R., Alawadi, A. H., Alsaalamy, A., Qasim, Q. A., & Sameni, F. (2024). The role of gut microbiota in human metabolism and inflammatory diseases: a focus on elderly individuals. *Annals of Microbiology 2023 74:1*, 74(1), 1–24.
- Khaliq, A. M., Erdogan, C., Kurt, Z., Turgut, S. S., Grunvald, M. W., Rand, T., Khare, S., Borgia, J. A., Hayden, D. M., Pappas, S. G., Govekar, H. R., Kam, A. E., Reiser, J., Turaga, K., Radovich, M., Zang, Y., Qiu, Y., Liu, Y., Fishel, M. L., ... Masood, A. (2022). Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biology*, 23(1).
- Kim, S. H., & Lim, Y. J. (2022). The role of microbiome in colorectal carcinogenesis and its clinical potential as a target for cancer treatment. *Intestinal Research*, 20(1), 31–42.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012 9:4, 9(4), 357–359.
- Lauby-Secretan, B., Vilahur, N., Bianchini, F., Guha, N., & Straif, K. (2018). *The IARC Perspective on Colorectal Cancer Screening*.

- Lee, J. K., Liles, E. G., Bent, S., Levin, T. R., & Corley, D. A. (2014). Accuracy of Fecal Immunochemical Tests for Colorectal Cancer. *Annals of Internal Medicine*, 160(3), 171–181.
- Lee, J. W. J., Plichta, D., Hogstrom, L., Borren, N. Z., Lau, H., Gregory, S. M., Tan, W., Khalili, H., Clish, C., Vlamakis, H., Xavier, R. J., & Ananthakrishnan, A. N. (2021). Multi-omics reveal microbial determinants impacting responses to biologic therapies in inflammatory bowel disease. *Cell Host & Microbe*, 29(8), 1294-1304.e4.
- Lin, D., Fan, W., Zhang, R., Zhao, E., Li, P., Zhou, W., Peng, J., & Li, L. (2021). Molecular subtype identification and prognosis stratification by a metabolism-related gene expression signature in colorectal cancer. *Journal of Translational Medicine*, 19(1).
- Lin, H., & Peddada, S. Das. (2020). Analysis of compositions of microbiomes with bias correction. *Nature Communications* 2020 11:1, 11(1), 1–11.
- Liu, C., Li, Z., Ding, J., Zhen, H., Fang, M., & Nie, C. (2021). Species-Level Analysis of the Human Gut Microbiome Shows Antibiotic Resistance Genes Associated With Colorectal Cancer. *Frontiers in Microbiology*, 12.
- Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., Qin, N., & Zhang, H. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Scientific Reports*, 6.
- Liu, X., Mao, B., Gu, J., Wu, J., Cui, S., Wang, G., Zhao, J., Zhang, H., & Chen, W. (2021). *Blautia*—a new functional genus with potential probiotic properties? In *Gut Microbes* (Vol. 13, Issue 1, pp. 1–21). Bellwether Publishing, Ltd.
- Liu, Y., Li, X., Yang, Y., Liu, Y., Wang, S., Ji, B., & Wei, Y. (2021). Exploring Gut Microbiota in Patients with Colorectal Disease Based on 16S rRNA Gene Amplicon and Shallow Metagenomic Sequencing. *Frontiers in Molecular Biosciences*, 8.
- Liu, Y., Zhang, S., Zhou, W., Hu, D., Xu, H., & Ji, G. (2022). Secondary Bile Acids and Tumorigenesis in Colorectal Cancer. *Frontiers in Oncology*, 12, 813745.
- Lv, M., Zhang, J., Deng, J., Hu, J., Zhong, Q., Su, M., Lin, D., Xu, T., Bai, X., Li, J., & Guo, X. (2023). Analysis of the relationship between the gut microbiota enterotypes and colorectal adenoma. *Frontiers in Microbiology*, 14, 1097892.
- Martinez-Medina, J. N., Flores-Lopez, R., López-Contreras, B. E., Villamil-Ramirez, H., Guzman-Muñoz, D., Macias-Kauffer, L. R., León-Mimila, P., Granados-Portillo, O., Del-Rio-navarro, B. E., Gómez-Perez, F. J., Aguilar-Salinas, C. A., Torres, N., Tovar, A. R., Canizales-Quinteros, S., & Moran-Ramos, S. (2021). Effect of Gut Microbial Enterotypes on the Association between Habitual Dietary Fiber Intake and Insulin Resistance Markers in Mexican Children and Adults. *Nutrients*, 13(11).
- McQuade, J. L., Daniel, C. R., Helmink, B. A., & Wargo, J. A. (2019). Modulating the microbiome to improve therapeutic response in cancer. In *The Lancet Oncology* (Vol. 20, Issue 2, pp. e77–e91). Lancet Publishing Group.

- Mizutani, S., Yamada, T., & Yachida, S. (2020). Significance of the gut microbiome in multistep colorectal carcinogenesis. *Cancer Science*, 111(3), 766–773.
- Mohajeri, M. H., Brummer, R. J. M., Rastall, R. A., Weersma, R. K., Harmsen, H. J. M., Faas, M., & Eggersdorfer, M. (2018). The role of the microbiome for human health: from basic science to clinical applications. *European Journal of Nutrition*, 57, 1–14.
- Nagata, N., Nishijima, S., Kojima, Y., Hisada, Y., Imbe, K., Miyoshi-Akiyama, T., Suda, W., Kimura, M., Aoki, R., Sekine, K., Ohsugi, M., Miki, K., Osawa, T., Ueki, K., Oka, S., Mizokami, M., Kartal, E., Schmidt, T. S. B., Molina-Montes, E., ... Kawai, T. (2022). Metagenomic Identification of Microbial Signatures Predicting Pancreatic Cancer From a Multinational Study. *Gastroenterology*, 163(1), 222–238.
- NIH Human Microbiome Portfolio Analysis Team. (2019). A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. In *Microbiome* (Vol. 7, Issue 1, p. 31). NLM (Medline).
- Oh, B., Boyle, F., Pavlakis, N., Clarke, S., Eade, T., Hruby, G., Lamoury, G., Carroll, S., Morgia, M., Kneebone, A., Stevens, M., Liu, W., Corless, B., Molloy, M., Kong, B., Libermann, T., Rosenthal, D., & Back, M. (2021). The Gut Microbiome and Cancer Immunotherapy: Can We Use the Gut Microbiome as a Predictive Biomarker for Clinical Response in Cancer Immunotherapy? *Cancers 2021*, Vol. 13, Page 4824, 13(19), 4824.
- Olovo, C. V., Huang, X., Zheng, X., & Xu, M. (2021). Faecal microbial biomarkers in early diagnosis of colorectal cancer. *Journal of Cellular and Molecular Medicine*, 25(23), 10783–10797.
- Parker, B. J., Wearsch, P. A., Veloo, A. C. M., & Rodriguez-Palacios, A. (2020). The Genus *Alistipes*: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. *Frontiers in Immunology*, 11, 906.
- Pericleous, M., Mandair, D., & Caplin, M. E. (2013). Diet and supplements and their impact on colorectal cancer. *Journal of Gastrointestinal Oncology*, 4(4), 409.
- Peterson, D., Bonham, K. S., Rowland, S., Pattanayak, C. W., Klepac-Ceraj, V., Deoni, S. C. L., D'Sa, V., Bruchhage, M., Volpe, A., Beauchemin, J., Wallace, C., Rogers, J., Cano, R., Fernandes, J., Walsh, E., Rhodes, B., Huentelman, M., Lewis, C., De Both, M. D., ... Braun, J. (2021). Comparative Analysis of 16S rRNA Gene and Metagenome Sequencing in Pediatric Gut Microbiomes. *Frontiers in Microbiology*, 12.
- Ponziani, F. R., Bhoori, S., Castelli, C., Putignani, L., Rivoltini, L., Del Chierico, F., Sanguinetti, M., Morelli, D., Paroni Sterbini, F., Petito, V., Reddel, S., Calvani, R., Camisaschi, C., Picca, A., Tuccitto, A., Gasbarrini, A., Pompili, M., & Mazzaferro, V. (2019). Hepatocellular Carcinoma Is Associated With Gut Microbiota Profile and Inflammation in Nonalcoholic Fatty Liver Disease. *Hepatology (Baltimore, Md.)*, 69(1), 107–120.

- Qi, Y., Jiang, C., Cheng, J., Krausz, K. W., Li, T., Ferrell, J. M., Gonzalez, F. J., & Chiang, J. Y. L. (2015). Bile acid signaling in lipid metabolism: metabolomic and lipidomic analysis of lipid and bile acid markers linked to anti-obesity and anti-diabetes in mice. *Biochimica et Biophysica Acta*, 1851(1), 19–29.
- Ramon, E., Obón-Santacana, M., Khannous-Lleiffe, O., Saus, E., Gabaldón, T., Guinó, E., Bars-Cortina, D., Ibáñez-Sanz, G., Rodríguez-Alonso, L., Mata, A., García-Rodríguez, A., & Moreno, V. (2024). Performance of a Shotgun Prediction Model for Colorectal Cancer When Using 16S rRNA Sequencing Data. *International Journal of Molecular Sciences*, 25(2).
- Rideout, J. R., Chase, J. H., Bolyen, E., Ackermann, G., González, A., Knight, R., & Caporaso, G. G. (2016). Keemei: Cloud-based validation of tabular bioinformatics file formats in Google Sheets. *GigaScience*, 5(1).
- Ridlon, J. M., & Hylemon, P. B. (2012). Identification and characterization of two bile acid coenzyme A transferases from *Clostridium scindens*, a bile acid 7 α -dehydroxylating intestinal bacterium. *Journal of Lipid Research*, 53(1), 66–76.
- Salillas, S., & Sancho, J. (2020). Flavodoxins as Novel Therapeutic Targets against *Helicobacter pylori* and Other Gastric Pathogens. *International Journal of Molecular Sciences* 2020, Vol. 21, Page 1881, 21(5), 1881.
- Saus, E., Iraola-Guzmán, S., Willis, J. R., Brunet-Vega, A., & Gabaldón, T. (2019). Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential. In *Molecular Aspects of Medicine* (Vol. 69, pp. 93–106). Elsevier Ltd.
- Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6), R60.
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 2012 9:8, 9(8), 811–814.
- Shen, X., Li, J., Li, J., Zhang, Y., Li, X., Cui, Y., Gao, Q., Chen, X., Chen, Y., & Fang, J. Y. (2021). Fecal Enterotoxigenic *Bacteroides fragilis*–*Peptostreptococcus stomatis*–*Parvimonas micra* Biomarker for Noninvasive Diagnosis and Prognosis of Colorectal Laterally Spreading Tumor. *Frontiers in Oncology*, 11, 661048.
- Shkoporov, A. N., Chaplin, A. V., Khokhlova, E. V., Shcherbakova, V. A., Motuzova, O. V., Bozhenko, V. K., Kafarskaia, L. I., & Efimov, B. A. (2015). *Alistipes inops* sp. nov. and *Coprobacter secundus* sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, 65(12), 4580–4588.
- Shuwen, H., Yinhang, W., Xingming, Z., Jing, Z., Jinxin, L., Wei, W., & Kefeng, D. (2022). Using whole-genome sequencing (WGS) to plot colorectal cancer-related gut microbiota in a population with varied geography. *Gut Pathogens*, 14(1).
- Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1), 7–34.

- Singh, J., Rivenson, A., Tomita, M., Shimamura, S., Ishibashi, N., & Reddy, B. S. (1997). *Bifidobacterium longum*, a lactic acid-producing intestinal bacterium inhibits colon cancer and modulates the intermediate biomarkers of colon carcinogenesis. *Carcinogenesis*, 18(4), 833–841.
- Sørli, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., & Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), 10869–10874.
- Sørli, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A. L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14), 8418–8423.
- Strum, W. B. (2016). Colorectal Adenomas. *New England Journal of Medicine*, 374(11), 1065–1075.
- Su, X., Gao, Y., & Yang, R. (2022). Gut Microbiota-Derived Tryptophan Metabolites Maintain Gut and Systemic Homeostasis. *Cells*, 11(15).
- Sullivan, Å. (2001). Effect of antimicrobial agents on the ecological balance of human microflora. In *Lancet Infectious Diseases* (Vol. 1, Issue 2, pp. 101–114).
- Sun, T., Liu, S., Zhou, Y., Yao, Z., Zhang, D., Cao, S., Oncotarget, Z. W.-, & 2017, undefined. (n.d.). Evolutionary biologic changes of gut microbiota in an 'adenoma-carcinoma sequence' mouse colorectal cancer model induced by 1, 2-Dimethylhydrazine. *Ncbi.Nlm.Nih.Gov* T Sun, S Liu, Y Zhou, Z Yao, D Zhang, S Cao, Z Wei, B Tan, Y Li, Z Lian, S Wang Oncotarget, 2017•ncbi.Nlm.Nih.Gov. Retrieved January 29, 2024, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5352133/>.
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932.
- Takahashi, K., Nishida, A., Fujimoto, T., Fujii, M., Shioya, M., Imaeda, H., Inatomi, O., Bamba, S., Andoh, A., & Sugimoto, M. (2016). Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease. *Digestion*, 93(1), 59–65.
- Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., & Letellier, E. (2020). Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? *Trends in Microbiology*, 28(5), 401–423.
- The decoy genome*. (n.d.). Retrieved January 29, 2024, from <https://www.cureffi.org/2013/02/01/the-decoy-genome/>

- Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., ... Segata, N. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, 25(4), 667–678.
- Tian, Y., Xu, Q., Sun, L., Ye, Y., & Ji, G. (2018). Short-chain fatty acids administration is protective in colitis-associated colorectal cancer development. *The Journal of Nutritional Biochemistry*, 57, 103–109.
- Tilg, H., Adolph, T. E., Gerner, R. R., & Moschen, A. R. (2018). The Intestinal Microbiota in Colorectal Cancer. *Cancer Cell*, 33(6), 954–964.
- Torres-Maravilla, E., Boucard, A. S., Mohseni, A. H., Taghinezhad-S, S., Cortes-Perez, N. G., & Bermúdez-Humarán, L. G. (2021). Role of Gut Microbiota and Probiotics in Colorectal Cancer: Onset and Progression. *Microorganisms*, 9(5).
- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10), 902–903.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. In *Nature* (Vol. 449, Issue 7164, pp. 804–810). Nature Publishing Group.
- Uppakarn, K., Bangpanwimon, K., Hongpattarakere, T., Wanitsuwan, W., & Article, R. (2021). *Comparison of the Human Gut Microbiota between Normal Control Subjects and Patients with Colonic Polyps and Colorectal Cancer*. <https://doi.org/10.21203/rs.3.rs-1172479/v1>.
- Vacante, M., Ciuni, R., Basile, F., & Biondi, A. (2020). Gut Microbiota and Colorectal Cancer Development: A Closer Look to the Adenoma-Carcinoma Sequence. *Biomedicines*, 8(11), 1–19.
- Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., Schiweck, C., Kurilshikov, A., Joossens, M., Wijnemga, C., Claes, S., Van Oudenhove, L., Zhernakova, A., Vieira-Silva, S., & Raes, J. (2019). The neuroactive potential of the human gut microbiota in quality of life and depression. *Nature Microbiology* 2019 4:4, 4(4), 623–632.
- Vieira-Silva, S., Sabino, J., Valles-Colomer, M., Falony, G., Kathagen, G., Caenepeel, C., Cleynen, I., van der Merwe, S., Vermeire, S., & Raes, J. (2019). Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nature Microbiology* 2019 4:11, 4(11), 1826–1831.
- Vivarelli, S., Salemi, R., Candido, S., Falzone, L., Santagati, M., Stefani, S., Torino, F., Banna, G. L., Tonini, G., & Libra, M. (2019). Gut microbiota and cancer: From pathogenesis to therapy. In *Cancers* (Vol. 11, Issue 1). MDPI AG.

- Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., Goedert, J. J., Shi, J., Bork, P., & Sinha, R. (2016). Colorectal cancer and the human gut microbiome: Reproducibility with whole-genome shotgun sequencing. *PLoS ONE*, 11(5).
- Wang, Y., Zhang, Y., Qian, Y., Xie, Y. H., Jiang, S. S., Kang, Z. R., Chen, Y. X., Chen, Z. F., & Fang, J. Y. (2021). Alterations in the oral and gut microbiome of colorectal cancer patients and association with host clinical factors. *International Journal of Cancer*, 149(4), 925–935.
- Wang, Z., Dan, W., Zhang, N., Fang, J., & Yang, Y. (2023). Colorectal cancer and gut microbiota studies in China. *Gut Microbes*, 15(1).
- Warnke-Sommer, J., & Ali, H. (2017). *Graph-based Analysis of Genetic Features Associated with Mobile Elements in Crohn's Disease and Healthy Gut Microbiomes*.
- Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L. P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A. M., Manghi, P., Gandini, S., ... Zeller, G. (2019a). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25(4), 679–689.
- Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L. P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A. M., Manghi, P., Gandini, S., ... Zeller, G. (2019b). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25(4), 679–689.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., MacIejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, Di., ... Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1), D1074–D1082.
- Wong, S. H., Kwong, T. N. Y., Wu, C. Y., & Yu, J. (2019). Clinical applications of gut microbiota in cancer biology. In *Seminars in Cancer Biology* (Vol. 55, pp. 28–36). Academic Press.
- Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., & Lewis, J. D. (2011a). Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.)*, 334(6052), 105–108.
- Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., & Lewis, J. D. (2011b). Linking

long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.)*, 334(6052), 105–108.

Wu, X., Zhang, T., Zhang, T. S., & Park, S. (2024). The impact of gut microbiome enterotypes on ulcerative colitis: identifying key bacterial species and revealing species co-occurrence networks using machine learning. *Gut Microbes*, 16(1).

Wu, Y., Jiao, N., Zhu, R., Zhang, Y., Wu, D., Wang, A. J., Fang, S., Tao, L., Li, Y., Cheng, S., He, X., Lan, P., Tian, C., Liu, N. N., & Zhu, L. (2021). Identification of microbial markers across populations in early detection of colorectal cancer. *Nature Communications*, 12(1), 1.

Xu, S., Yin, W., Zhang, Y., Lv, Q., Yang, Y., & He, J. (2020). Foes or Friends? Bacteria Enriched in the Tumor Microenvironment of Colorectal Cancer. *Cancers*, 12(2).

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., Hosoda, F., Rokutan, H., Matsumoto, M., Takamaru, H., Yamada, M., Matsuda, T., Iwasaki, M., Yamaji, T., Yachida, T., ... Yamada, T. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature Medicine*, 25(6), 968–976.

Yamauchi, M., Morikawa, T., Kuchiba, A., Imamura, Y., Qian, Z. R., Nishihara, R., Liao, X., Waldron, L., Hoshida, Y., Huttenhower, C., Chan, A. T., Giovannucci, E., Fuchs, C., & Ogino, S. (2012). Assessment of Colorectal Cancer Molecular Features along Bowel Subsites Challenges the Conception of Distinct Dichotomy of Proximal vs. Distal Colorectum. *Gut*, 61(6), 847.

Yang, T. W., Lee, W. H., Tu, S. J., Huang, W. C., Chen, H. M., Sun, T. H., ... & Lin, C. C. (2019). Enterotype-based analysis of gut microbiota along the conventional adenoma-carcinoma colorectal cancer pathway. *Scientific reports*, 9(1), 10923.

Yang, Y., Du, L., Shi, D., Kong, C., Liu, J., Liu, G., Li, X., & Ma, Y. (2021). Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature Communications*, 12(1).

Young, C., Wood, H. M., Seshadri, R. A., Van Nang, P., Vaccaro, C., Melendez, L. C., Bose, M., Van Doi, M., Piñero, T. A., Valladares, C. T., Arguero, J., Balaguer, A. F., Thompson, K. N., Yan, Y., Huttenhower, C., & Quirke, P. (2021). The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. *Genome Medicine*, 13(1).

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Yi Liang, Q., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., Wang, X., Xu, X., Chen, N., Wu, W. K. K., Al-Aama, J., Nielsen, H. J., Kiilerich, P., Jensen, B. A. H., Yau, T. O., ... Wang, J. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1), 70–78.

- Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., & Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research*, 7(11), 1112–1121.
- Ze, X., Duncan, S. H., Louis, P., & Flint, H. J. (2012). *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *The ISME Journal*, 6(8), 1535–1543.
- Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D. R., Schneider, M. A., Schrotz-King, P., Tournigand, C., Tran Van Nhieu, J., Yamada, T., ... Bork, P. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11).
- Zhang, D., Jian, Y. P., Zhang, Y. N., Li, Y., Gu, L. T., Sun, H. H., Liu, M. Di, Zhou, H. L., Wang, Y. S., & Xu, Z. X. (2023). Short-chain fatty acids in diseases. *Cell Communication and Signaling 2023 21:1*, 21(1), 1–20.
- Zhu, L. Bin, Zhang, Y. C., Huang, H. H., & Lin, J. (2021). Prospects for clinical applications of butyrate-producing bacteria. *World Journal of Clinical Pediatrics*, 10(5), 84–92.
- Zitvogel, L., Daillère, R., Roberti, M. P., Routy, B., & Kroemer, G. (2017). Anticancer effects of the microbiome and its products. In *Nature Reviews Microbiology* (Vol. 15, Issue 8, pp. 465–478). Nature Publishing Group.
- Zmora, N., Zilberman-Schapira, G., Suez, J., Mor, U., Dori-Bachash, M., Bashiares, S., Kotler, E., Zur, M., Regev-Lehavi, D., Brik, R. B. Z., Federici, S., Cohen, Y., Linevsky, R., Rothschild, D., Moor, A. E., Ben-Moshe, S., Harmelin, A., Itzkovitz, S., Maharshak, N., ... Elinav, E. (2018). Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell*, 174(6), 1388-1405.e21.

BIOGRAPHY

Ünzile GUVEN GULHAN completed her B.S. degree in the Department of Biology and the Department of Genetics and Bioengineering at Fatih University, Istanbul, both in 2011. She completed her M.S. degree in the Department of Genetics and Bioengineering at Fatih University, Istanbul, in 2014. She is currently a Ph.D. student at the Systems Biology and Bioinformatics Program at Gebze Technical University, Kocaeli , with a focus on the identification of metagenomic markers related to colon cancer diagnosis. She has been working at PhiTech Bioinformatics as a bioinformatician since 2017. She is interested in the area of bioinformatics research and development area; NGS, genomics, metagenomics and human gut microbiota.

PUBLICATIONS AND PRESENTATIONS FROM THE THESIS

Gülhan, Ü. G., Nikerel, E., Çakır, T., Sevilgen, F. E., & Durmuş, S. (2024). Species-level identification of enterotype-specific microbial markers for colorectal cancer and adenoma. *Molecular Omics*.

APPENDICES

APPENDIX-A: Drug Molecules Targeting CRC-Associated Homologous Proteins in the DrugBank Database

ET 1, CRC associated protein	Homologous protein	Targeted drug	Group	Targeted organism
UniRef90_A8AMJ6_30S ribosomal protein S18	Elongation factor Tu, 30S ribosomal protein	2-METHYLTHIO-N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE	Experimental	<i>Thermus aquaticus</i> , <i>Thermus thermophilus</i>
UniRef90_P68768_Cytosolaminopeptidase	Cytosol aminopeptidase, Bacterial leucyl aminopeptidase	Leucine Phosphonic Acid	Experimental	Human, <i>Vibrio proteolyticus</i>
	Cytosol aminopeptidase	(2S)-3-[(R)-[(1S)-1-amino-3-phenylpropyl](hydroxy)phosphoryl]-2-benzylpropanoic acid		Human
		Zofenoprilat		
ET 2, CRC associated protein	Homologous protein	Targeted drug	Group	Targeted organism
UniRef90_B5CUJ8_Integrase core domain protein	Integrase, catalytic core	Pipendoxifene	Clinical trial	<i>Nitrosomonas europaea</i>
UniRef90_I3YM16_50S ribosomal protein L3	50S ribosomal protein L3	Retapamulin	Approved	<i>Streptococcus pyogenes</i> serotype M1
		Spiramycin		
UniRef90_D4IMJ2_30S ribosomal protein S10	30S ribosomal protein S10	Nitrofurantoin	Approved	<i>Escherichia coli</i> (strain K12)
	30S ribosomal protein S10, 16S ribosomal RNA	Paromomycin	Approved, Clinical trial	<i>Thermus thermophilus</i> , <i>Enteric bacteria</i> and other eubacteria
	Elongation factor Tu, 30S ribosomal protein	2-METHYLTHIO-N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE	Experimental	<i>Thermus aquaticus</i> , <i>Thermus thermophilus</i>

ET 3, CRC associated protein	Homologous protein	Targeted drug	Group	Targeted organism
UniRef90_P30899_Beta_lactamase	Beta-lactamase	Avibactam	Approved	<i>Escherichia coli</i> , <i>Enterobacter cloacae</i> , <i>Klebsiella pneumoniae</i> , <i>Pseudomonas aeruginosa</i>
		Vaborbactam	Approved, Clinical trial	<i>Enterobacter cloacae</i> , <i>Klebsiella pneumoniae</i> , <i>Escherichia coli</i>
		Relebactam		<i>Escherichia coli</i> , <i>Klebsiella pneumoniae</i> , <i>Enterobacter cloacae</i>
	Penicillin-binding protei	Imipenem	Approved	<i>Escherichia coli</i> (strain K12), <i>Bacillus subtilis</i> (strain 168), <i>Staphylococcus aureus</i>
UniRef90_Q5L8Z8_Malate dehydrogenase	Fumarylacetoacetase, Malate dehydrogenase, Fumarate reductase flavoprotein subunit	Fumaric acid	Experimental	<i>Human</i> , <i>Chloroflexus aurantiacus</i> , <i>Shewanella frigidimarina</i> , <i>Shewanella oneidensis</i> (strain MR-1)
	Fumarylacetoacetase, Malate dehydrogenase, Fumarate reductase flavoprotein subunit	3-hydroxyisoxazole-4-carboxylic acid		<i>Plasmodium falciparum</i>
	L-lactate dehydrogenase	4-Hydroxy-1,2,5-oxadiazole-3-carboxylic acid		<i>Plasmodium falciparum</i> , Human
	L-lactate dehydrogenase	4-oxo-4,5-dihydro-1,2,5-thiadiazole-3-carboxylic acid		<i>Plasmodium falciparum</i> , Human, <i>Thermotoga maritima</i> , <i>Escherichia coli</i> , <i>Lactobacillus helveticus</i>
	L-lactate dehydrogenase	Oxamic Acid		<i>Plasmodium falciparum</i> , <i>Xanthomonas</i>

				<i>campestris</i> pv. <i>Campestris</i>
	L-lactate dehydrogenase, Formate acetyltransferase	Naphthalene-2,6-disulfonic acid		<i>Plasmodium falciparum</i>
	L-lactate dehydrogenase, Bacterioferritin comigratory protein	3,7-DIHYDROXYNAPHTHALENE-2-CARBOXYLIC ACID		Human, <i>Plasmodium falciparum</i>
UniRef90_Q8A485_30S ribosomal protein S171	Elongation factor Tu, 30S ribosomal protein	2-METHYLTHIO-N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE	Experimental	<i>Thermus aquaticus</i> , <i>Thermus thermophilus</i>
UniRef90_R5K610_Adenylosuccinate lyase	Fumarate hydratase class II	Pyromellitic Acid	Experimental	<i>Escherichia coli</i> (strain K12)
		3-Trimethylsilylsuccinic Acid		

APPENDIX-B: Drug Molecules Targeting ADN-Associated Homologous Proteins in the DrugBank Database

ET 1, ADN associated protein	Homologous protein	Targeted drug	Group	Targeted organism
UniRef90_D4KJY8_Response regulators consisting of a CheY_like receiver domain and a winged helix DNA binding domain*	Adenylate cyclase	Adenosine-5'-Rp-Alpha-Thio-Triphosphate	Experimental	<i>Arthrospira platensis</i>
		2-Hydroxyestradiol		
	Beta-phosphoglucosyltransferase, Nitrogen regulation protein NR(I)	Phosphoaspartate		<i>Lactococcus lactis subsp. lactis (strain IL1403), Salmonella typhimurium</i>
UniRef90_C4ZBS3_30S ribosomal protein S19	Elongation factor Tu, 30S ribosomal protein	2-METHYLTHIO-N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE	Experimental	<i>Thermus aquaticus, Thermus thermophilus</i>
	30S ribosomal protein	Tigecycline	Approved	<i>Escherichia coli (strain K12)</i>
		Chlortetracycline	Approved, Clinical trial	
UniRef90_C4Z2R6_30S ribosomal protein S12	30S ribosomal protein	Amikacin	Approved	<i>Escherichia coli (strain K12)</i>
		Arbekacin	Experimental, Clinical trial	
		Spectinomycin	Approved, Clinical trial	
	30S,16S ribosomal protein	Netilmicin	Approved, Clinical trial	<i>Escherichia coli (strain K12), Enteric bacteria and other eubacteria</i>
		Kanamycin		
	30S ribosomal protein, 16S-23S rRNA, NH(3)-dependent NAD(+) synthetase, Dihydrofolate reductase	Gentamicin	Approved	<i>Escherichia coli (strain K12), Enteric bacteria and other eubacteria, Bacillus subtilis (strain 168), Human</i>

	Elongation factor Tu, 30S ribosomal protein	2-METHYLTHIO-N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE	Experimental	<i>Thermus aquaticus</i> , <i>Thermus thermophilus</i>	
UniRef90_R6BEM2_L_aspartate oxidase	Succinate dehydrogenase subunits	Atpenin A5	Experimental	<i>Escherichia coli</i> (strain K12)	
		2-[1-METHYLHEXYL]-4,6-DINITROPHENOL			
	Fumarate reductase subunit	Thiabendazole	Approved		<i>Shewanella frigidimarina</i>
		2-[1-(4-CHLOROPHENYL)-ETHYL]-4,6-DINITROPHENOL	Experimental		
		Malate Like Intermediate			
		Citraconic acid			
	2,3-Dimethyl-1,4-naphthoquinone		<i>Wolinella succinogenes</i>		
UniRef90_D4LR50_Monosaccharide ABC transporter ATP_binding protein_CUT2 family_TC3_A_1_2_	HIV-1 protease	Saquinavir	Approved, Clinical trial	<i>Human immunodeficiency virus 1</i>	
		Indinavir	Approved		
		Amprenavir	Approved, Clinical trial		
		Nelfinavir	Approved		
	Polymerase acidic protein	Baloxavir marboxil	Approved, Clinical trial	<i>Influenza A virus</i>	
	16S ribosomal RNA	Doxycycline		<i>Enteric bacteria and other eubacteria</i>	
	DNA gyrase subunit A, DNA topoisomerase 4 subunit A	Delafloxacin	Approved	<i>Escherichia coli</i> (strain K12)	
	Dihydrofolate reductase	Trimethoprim			
	DNA topoisomerase	Novobiocin			<i>Staphylococcus aureus</i>

	1, DNA gyrase subunit B			
UniRef90_D4LM M9_50S ribosomal protein L3	50S ribosomal protein L3	Spiramycin Retapamulin	Approved	<i>Streptococcus pyogenes serotype M1</i>
ET 2, ADN associated protein	Homologous protein	Targeted drug	Group	Targeted organism
UniRef90_P0AEP 2_Galactose_prot on symporter	Aspartate aminotransferase, Glutamate decarboxylase alpha, Branched-chain-amino-acid aminotransferase	Glutaric Acid	Experimental	<i>Escherichia coli (strain K12)</i>
	HIV-1 protease	Saquinavir	Approved, Clinical trial	<i>Human immunodeficiency virus 1</i>
		Indinavir Nelfinavir	Approved	
UniRef90_A8AMJ 6_30S ribosomal protein S18	Elongation factor Tu, 30S ribosomal protein	2-METHYLTHIO-N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE	Experimental	<i>Thermus aquaticus, Thermus thermophilus</i>
ET 3, ADN-associated protein	Homologous protein	Targeted drug	Group	Targeted organism
UniRef90_R6U0H 8_Transcriptional regulator AraC family	Arabinose operon regulatory protein, L-arabinose-binding periplasmic protein	beta-D-fucose	Experimental	<i>Escherichia coli (strain K12)</i>
	L-arabinose-binding periplasmic protein, Aldose 1-epimerase	Alpha-L-Arabinose		<i>Escherichia coli (strain K12), Lactococcus lactis</i>

* common in all ETs