MICROBIOME ANALYSIS IN COLORECTAL CANCER

by

Ezzat Dadkhah

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Biosciences

Committee:

_____ Dr. Patrick Gillevet, Dissertation Director

_____ Dr. Ancha Baranova, Committee Member

_____ Dr. Donald Seto, Committee Member

_____ Dr. James Goedert, Committee Member

_____ Dr. Iosif Vaisman, Acting Director, School of Systems Biology

_____ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_____ Dr. Peggy Agouris, Dean, College of Science

Date: _____ Fall Semester 2017
George Mason University
Fairfax, VA

Microbiome Analysis in Colorectal Cancer

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Ezzat Dadkhah
Master of Science
Islamic Azad University, 2009

Director: Patrick Gillevet, Professor
Department of Biosciences

Fall Semester 2017
George Mason University
Fairfax, VA

DEDICATION

This dissertation dedicated to my mom & dad, my husband, Amir, my lovely little son, Ryan, and my sister, Katayoon.

ACKNOWLEDGEMENTS

I would like to thank the George Mason University, School of Systems Biology to give me the opportunity to train in an excellent professional and scientific environment, where I learned much from the professors, students, and staff. They are all amazing, and each of them helped me to become a better person than I was before attending this program.

I would like to thank Dr. Patrick Gillevet for all support and patience through this work. His exceptional training and advice provided a great opportunity for me to increase my knowledge in biology and bioinformatics. I appreciate Dr. Ancha Baranova, Dr. Don Seto and Dr. James Goedert for accepting to be a member of my committee and offer their valuable advice and guidance.

I owe my deep gratitude to the individuals who have participated in this study and many other people involved in this work either by recruiting subjects, taking samples, organizing sample storage, documenting clinical data, and labratoary works. Without this cooperation, this project would not have been possible.

Last but not least I am grateful to my family and friends. My husband, Amir, thanks for being at my side, even in difficult times, your kind words were always motivating to me. Thanks for your patience during these years and thank you for being there. And my parents, thank you for devoting your unconditional love. My sister, Katayoon, you were amazing. There are some helps that can just be received from a loving sister and you've given it thoroughly. My truly friend, Ekaterina Marakasova, I am grateful for being such an amazing person. You were a role model of srength and optimism for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Adenoma	Ad
Area under curve	AUC
Classification accuracy	CA
Colorectal cancer	CRC
False discovery rate	FDR
Fecal immunochemical test	FIT
Flexible sigmoidoscopy	FSIG
Gastrointestinal	GI
Guaiac fecal occult blood test	gFOBT
Inflammatory bowel disease	IBD
Next generation sequencing	NGS
Kyoto Encyclopedia of Genes and Genomes	KEGG
Operational Taxonomic Unit	OTU
Phylogenetic Investigation of Communities by Reconstruction of Unobserved States	PICRUSt
Polymerase chain reaction	PCR
Principal coordinates analysis	PCoA
Quantitative PCR	QPCR
Ribonucleic acid	RNA
Ribosomal Database Project	RDP
Ribosomal RNA	rRNA
Sensitivity	Sens
Short chain fatty acid	SCFA
Specificity	Spec
Support vector machines	SVM
True positive rate	TPR
Unique fraction	UniFrac

ABSTRACT

MICROBIOME ANALYSIS IN COLORECTAL CANCER

Ezzat Dadkhah, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. Patrick Gillevet

Colorectal cancer (CRC) results from a complex interplay between genes and the environment. Recent studies have focused on the gut microbial population (the microbiota) and its aggregate genome (the microbiome) as one of the environmental players in colorectal tumorigenesis. High-throughput sequencing techniques have added a new dimension to the mining of gut microbiome for biomarkers of CRC and therapeutic targets. Current approaches to microbiome analysis include quantifying the relative abundancies and diversities of microbial populations along with the identification of disease-specific biomarkers.

In this project, the 16S rRNA sequences of bacteria present in stool samples of patients with CRC, pre-cancerous adenomatous polyps, and non-cancer controls were analyzed using three different operational taxonomic units (OTUs) identification techniques - UPARSE, UPGMA, and UCLUST. UPARSE was the fastest algorithm and identified the lowest number of OTUs while UPGMA required the largest amount of

memory. UCLUST was the slowest and identified the highest number of OTUs. The patterns of alpha diversity (diversity within a sample) and beta diversity (diversity between samples) obtained by each of these algorithms were not substantially different.

In this dissertation, we report the analysis of samples collected from subjects that have undergone routine colonoscopy to detect the presence of polyp(s). Various nonparametric statistics and classification techniques were utilized to identify the microbiota characteristics capable of discriminating between disease states and from healthy colon. OTUs significantly different in their relative abundance between subjects with polyp (polyp-Y group) and without polyp (polyp-N group) were used to build classifying predictors for the presence or absence of polyps.

The predictive power to discriminate between polyp-N and polyp-Y groups was highest when the model was built using OTUs preselected for statistically significant differences in their relative abundance. In conclusion, we showed that 16S rRNA microbiome analysis could be utilized to generate OTU abundance-based feature sets for further development into the predictive models. Eventually, these models will improve the power of CRC diagnostics and aid in defining the dynamic interface between the gut and residing microbiota.

1 AIM OF STUDY

Colorectal cancer (CRC) has recently been reported to be the second leading cause of cancer death in the United States, and the number of deaths due to CRC are estimated to be about 50,000 annually (Siegel et al., 2017). There is an extensive, diverse microbial population in the gut which plays a significant role in the maintenance of health and immunity (Guarner & Malagelada, 2003; Salminen et al., 1998). Growing evidence suggests that the gut microbiome contributes to colorectal carcinogenesis (Keku et al., 2015). In particular, experiments where the stool samples of mice with colon cancer or stool of a human patient with CRC were transplanted to germ-free animals are indicative of that (Zackular et al., 2013; Baxter et al., 2014). Furthermore, some studies have linked the gut microbiome composition to the development of colon polyps and adenomas (Ahn et al., 2013; Keku et al., 2015; Dulal & Keku, 2014).

In this study, we investigate the gut microbiome differences in individuals who underwent a colonoscopy and were categorized into two groups of subjects, those with polyp (polyp-Y) and those without polyps (polyp-N). In addition to the polyp dataset analysis, we analyze raw 16S rRNA sequences from previously published studies using the three classic sequencing analysis approaches, UPARSE, UPGMA, and UCLUST, which are currently the most used OTU selection approaches of microbiome analysis. The aim of

using different approaches is to find out which would be optimal for analysis of these datasets.

In previous studies, changes in abundance for some of the bacterial taxa were seen when cancer or adenoma patients were compared to control subjects. These taxa included the phyla Bacteroidetes and Firmicutes; the genera *Proteobacteria, Fusobacterium, Blautia, Bifidobacterium, and Roseburia* as well as *Bacteroides spp.* (Chen et al., 2013; Brim et al., 2013; Goedert et al., 2015; Shen et al., 2010; Mira-Pascual et al., 2015; Nugent et al., 2014; Zackular et al., 2014). We hypothesize that at the Phylum level, there is a significant difference in the abundance of Firmicutes and Bacteroidetes OTUs in the subjects with polyps or cancer in both the public datasets and in our polyp dataset, and at the genus level, there is a significant difference in the abundance of *Fusobacterium*, *Bacteroides*, *Blautia, Bifidobacterium,* and *Roseburia* in subjects with cancer and polyps in both the public dataset and our polyp dataset.

Moreover, we aim to evaluate the performance of various statistical tests to improve the feature selection, followed by application of machine learning approaches to discriminate disease versus healthy state based on abundancies of various 16S rRNA sequences, ultimately producing models for the prediction of the disease. Thus, we propose that the preselection of features using statistical tests will improve the classification performance to assist in the prediction of the presence of the disease in naïve patients who have not had a colonoscopy based diagnosis of colon carcinoma or benign polyp.

Public health officials, doctors, and the public at large are still looking for reliable noninvasive screening methods for CRC. This study could be a step forward to utilize the microbiome as a suitable screening method for adenoma and CRC.

2 BACKGROUND

2.1 Colorectal cancer (CRC)

Colorectal cancer (CRC) is the third most common cancer and fourth most common cause of cancer death in the world (Ferlay et al., 2015). More than 1.2 million new cases of CRC occur annually, most of which are sporadic and result from an accumulation of mutations and epigenetic alterations in a variety of genes. The sequential accumulation of genetic mutations in tumor suppressors or oncogenes happens over time and causes a sequential transition of the normal mucosa to pre-malignant polyps, to adenoma, and eventually to CRC (Sears et al., 2014). This transition is often described as the "adenoma-carcinoma sequence" (Vogelstein et al., 2013).

Most colon neoplasms are adenocarcinomas that originate from epithelial cells of the mucosa. These adenocarcinomas usually arise from polyps (Stewart 2006). The two precursors of CRC are conventional adenomas and serrated polyps. The "conventional" pathway to CRC starts with adenomatous polyposis coli (APC) mutation, followed by chromosomal instability, and the hypermethylation of CpG islands associated with tumor-suppressor genes. About 60% of CRC cases arise from the conventional pathway. Serrated polyps are categorized into three subtypes: sessile serrated adenomas/polyps, traditional serrated adenomas, and hyperplastic polyps. Recent findings suggest that progression through the serrated pathway may be responsible for 20% to 30% of sporadic CRCs (Peters et al., 2016).

2.1.1 CRC risk factors

Colorectal malignancies are associated with both genetic and environmental risk factors. Family history is one of the well-known genetic risk factors for CRC. Positive first- and even second- and third-degree family histories are associated with an increased risk for CRC (Taylor et al., 2010). Other host and environmental risk factors associated with CRC include high body mass index (BMI), obesity, diabetes, polyps, consumption of red and processed meats, beer drinking (≥ 2 drinks/day), and smoking (Shaukat et al., 2017; Renehan et al., 2008; Larsson et al., 2005; Strum 2016; Larsson et al., 2006; Zhang et al., 2015; Giovannucci et al., 2002). On the other hand, physical and recreational activities, post-menopausal hormone therapy, as well as the consumption of milk, calcium, vitamin D, and aspirin are all negatively associated with CRC (Samad et al., 2005; Grodstein et al., 1999; Cho et al., 2004; Ma et al., 2011; Dube et al., 2007). Recently, it has been suggested that altered composition of the gut microbiome (termed "dysbiosis") could be another risk factor for CRC (Wang et al., 2017).

Three genetic models for the development of CRC are summarized in Figure 1 (Fearon, 2011). In the model shown in Figure 1a, CRC originates with adenomatous polyps. The order of the mutations is not always as represented here, but these mutations are strongly associated with specific stages of CRC progression.

The other two models are shown in Figure 1b in which CRCs originate from inherited (top model) and sporadic (bottom model) mismatch repair gene (MMR) defects that have been reported to be responsible for 15% of CRCs (Fearon, 2011).

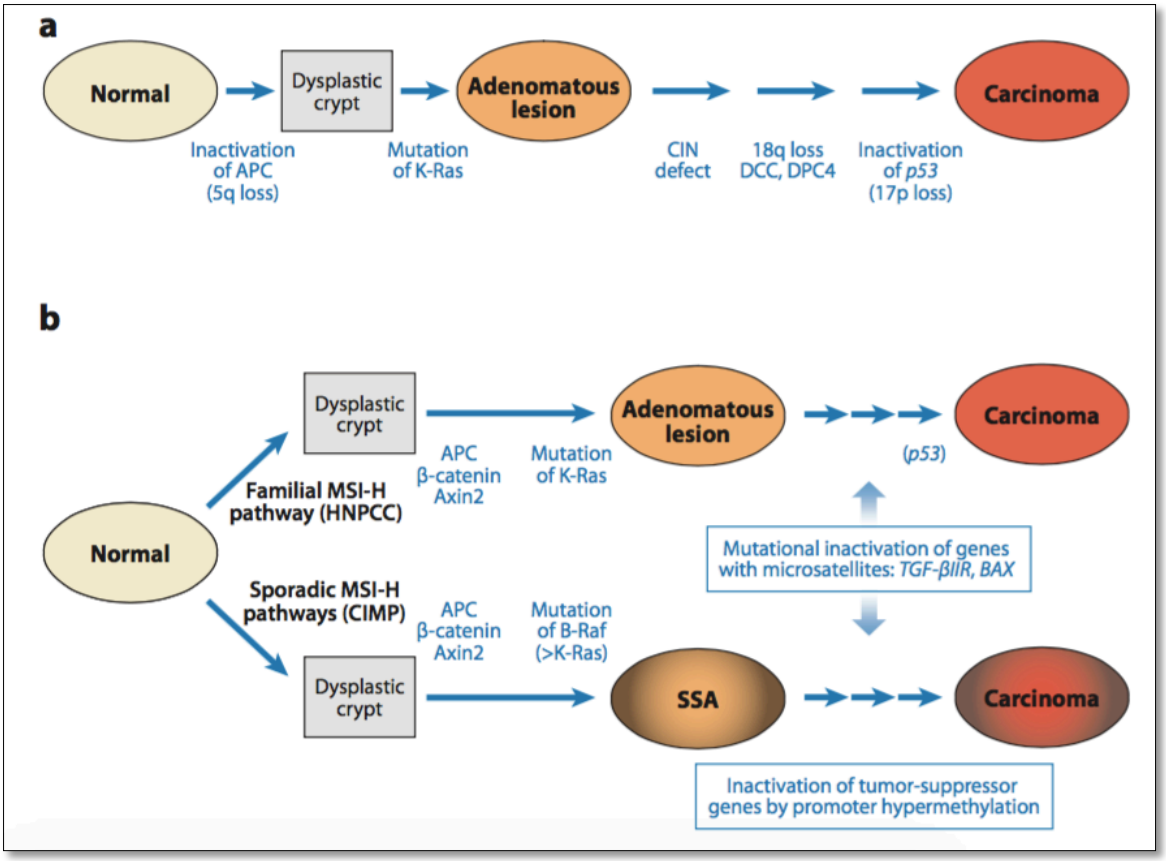


Figure 1 The three genetic models of colorectal cancer (CRC) progression.
(a) Most CRCs are developed as the result of APC mutations followed by subsequent genetic alterations. (*b*) The other two models start with somatic or epigenetic inactivation of mismatch repair (MMR) genes, which lead to high microsatellite instability (MSI-H). The following mutations ultimately promote cells to develop cancer phenotype. Abbreviations: APC, adenomatous polyposis coli; CIN, chromosome instability; DCC, deleted in colorectal carcinoma; DPC4, deleted in pancreatic carcinoma, locus 4; HNPCC, hereditary nonpolyposis colorectal cancer; CIMP, CpG island hypermethylation phenotype; SSA, sessile serrated adenomas. (Ref: Fearon, 2011).

The adenomatous polyposis coli (*APC*) gene encodes a tumor suppressor and is one of the earliest genes that mutates in CRC. Its product is a pleiotropic protein with versatile functions in apoptosis, cell cycle regulation, intercellular adhesion, and in the Wnt signaling pathway that modulates cell fate determination, cell migration, cell polarity, neural patterning, and organogenesis. Some of these *APC* mutations enable the cell to grow faster than normal. Other tumor suppressor genes that are commonly mutated in CRC are

the β-catenin gene (*CTNNB1*), the Deleted in CRC gene (*DCC*), and the P53 gene (*TP53*). Additionally, *KRAS* mutations support clonal cell division as this oncogene participates in G protein-mediated signal transduction and regulates cellular proliferation and differentiation (Fearon et al., 1990). These events can be followed by further mutations in *PIK3CA* (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha), *SMAD4*, *TP53*, and *BRAF* (B-Raf proto-oncogene, serine/threonine kinase) mutations, all of which potentiate malignancy. Tracking the specific pathological and genetic characteristics of adenomas can be helpful in predicting their potential to become malignant, as not all adenomas transform into malignancies (Mclean et al., 2011).

2.1.2 CRC screening

It is well accepted that increased participation in screening programs leads to earlier CRC diagnosis and helps to prevent the disease by the timely removal of polyps (Narayanan, 2014). Available screening methods for CRC include the guaiac fecal occult blood test (gFOBT), fecal immunochemical test (FIT), flexible sigmoidoscopy (FSIG), colonoscopy, colon-capsule endoscopy, and computed tomography (CT)-colonography. Typically, gFOBT and FIT techniques, which are the least expensive ones, are utilized by population-wide screening programs, while colonoscopy, CT-colonography, and FSIG are used in symptomatic patients (Narayanan, 2014). Additionally, a variety of molecular tests are available to diagnose CRC by utilizing blood, stool, or urine samples (Kuipers et al., 2013, Kuipers et al., 2014).

The guaiac fecal occult blood test (gFOBT): the blood vessels on the surface of colorectal polyps and cancers are fragile and easily break by the passage of stool. The small

amount of blood that is released from these vessels is rarely noticed by the subject but can be detected by the gFOBT test. This is a convenient and noninvasive method that can reduce the CRC-related mortality rate (Faivre et al., 2004). The gFOBT is a fecal test, which detects the presence of blood by the use of a paper impregnated with the chemical reagent guaiac. After a small amount of stool is smeared to a test card and hydro-peroxidase added, the heme will oxygenize the guaiac leading to the appearance of blue coloration. This standard qualitative (positive/negative) gFOBT test includes three paper cards that have two panels each and requiring sampling from three separate bowel movements. This test may be done with or without rehydration by adding water to the card before processing the sample. Rehydration method has higher sensitivity but results in larger amount of false-positive test results (Tinmouth et al., 2015).

Importantly, the gFOBT test cannot determine whether the bleeding is from the colon or other parts of the digestive tract such as the stomach. Therefore, positive tests are followed by a colonoscopy to determine if there is a polyp, tumor, or other problems such as hemorrhoids, ulcers, inflammatory bowel disease (colitis), or diverticulitis (i.e., small pouches in the colon wall). Additionally, the sensitivity of the test in detecting minor or non-bleeding CRC cases is poor. However, it has been shown that by repeating the fecal sampling and rerunning the test, the detection rate could improve from 9-12% to 52.6% (Lieberman et al., 2001). In fact, the timely administration of gFOBT may reduce CRC mortality rate by as much as 33% (Levin et al., 2008). Still, the sensitivity and specificity of a gFOBT are highly variable depending on the brand and the particular type of the test kit, specimen collection method, number of stool samples profiled per test, rehydration,

interpretation, and screening interval (Levin et al., 2008). The sensitivity of gFOBT for the presence of CRC varies from 12.9% to 79.4% and the specificity varies from 86.7% to 97.7% (Mandel et al., 1993; Hardcastle et al., 1996; Kronborg et al., 1996; Lieberman, 2009; Ahlquist et al., 1993; Ahlquist & Shuber, 2002; Allison et al., 1996; Imperiale et al., 2004). Therefore, a positive gFOBT test necessitates confirmation with other screening methods (Kuipers et al., 2013).

The fecal immunochemical test (FIT) has several advantages over the gFOBT. The FIT is quantitative; it detects intact hemoglobin rather than heme, and, therefore, is more specific. In this context, partially degraded hemoglobin can be present in the stool if bleeding originates in the upper gastrointestinal tract, while the presence of intact hemoglobin indicates lower gastrointestinal tract bleeding (Rockey, 1999). Thus, the major disadvantage of this method is its insensitivity in detecting proximal CRC as in this case the hemoglobin degrades before it reaches the distal colon and will not be present in the stool (Narayanan 2014).

In asymptomatic individuals with average risk for CRC, the overall sensitivity of FIT test is at 0.79 (95%CI: 0.69-0.86), and the specificity is at 0.94 (95%CI: 0.92-0.95) for FIT test (Allison et al., 1996; Lee et al., 2014; Allison et al., 2007; Cheng et al., 2002; Nakama et al., 1999; Nakama et al., 1996; Parra-Blanco et al., 2010; Chiu et al., 2013; Chiang et al., 2011; Sohn et al., 2005; Levi et al., 2011; Levi et al., 2007; Morikawa et al., 2005; Launoy et al., 2005; Itoh et al., 1996; Nakazato et al., 2006; Park et al., 2010; de Wijkerslooth et al., 2012; Brenner & Tao, 2013). Due to cost-efficiency of the FIT test,

this approach is the most commonly-used method for CRC screening (Song & Yu-Min, 2016).

The flexible sigmoidoscopy (FSIG): the FSIG is the endoscopic examination of the lower one-third of the colon lumen. The FSIG procedure is administered by trained non-physicians, requires less preparation for both patients and examiners than colonoscopy, and does not require sedation. FSIG is also less expensive than a colonoscopy and causes fewer complications. Case-control studies showed that FSIG leads to a 60-80% reduction in colon cancer mortality (Selby et al., 1992; Newcomb et al., 1992). The protective effect lasts from 5 to 10 years depending on the skill of the endoscopist, the quality of examination, and results of other annual screening tests such as gFOBT and FIT. The main drawback of FSIG is that the examined area is limited to the rectum, sigmoid and descending colon.

Colonoscopy: This approach is the best method for early detection and prevention of CRC as the entire colon is examined. If performed by experienced professionals, the sensitivity and specificity to diagnose advanced adenomas and cancer are as high as 100% (Garborg et al., 2013).

Colonoscopy is often performed as a follow-up to other positive CRC screening tests such as gFOBT and FIT, but randomized trials for colonoscopy screening are lacking. However, population-based case-control studies performed in Canada, Germany, and the United States, as well as a follow-up study of the National Polyp Study cohort suggest that colonoscopy may have, in fact, already reduced CRC incidence by 67-77% and CRC mortality rate by 31-65% (Baxter et al., 2009; Brenner et al., 2011; Kahi et al., 2009; Zauber

et al., 2012; Garborg et al., 2013). However, colonoscopy is invasive, time-consuming, expensive, associated with some complications and some discomfort, and requires a significant bowel preparation. Therefore, it is not widely accepted as a screening approach (Garborg et al., 2013).

Molecular markers: Molecular genetic markers are being developed for CRC screening from either fecal samples or the blood (Narayanan et al., 2014). Adenoma and carcinoma cells shed DNA into the large bowel lumen, and this DNA is then excreted in the stool where it remains relatively stable. Typically, human cells in the fecal samples undergo evaluation for the presence of DNA alterations previously identified as contributors to the adenoma-carcinoma sequence of colorectal carcinogenesis. To improve sensitivity, a panel of DNA alterations may be analyzed in a multiplexed test (Levin et al., 2008). In one study, a multi-target DNA test that detects mutations in *KRAS* (Kirsten rat sarcoma), aberrant methylated *NDRG4* (N-Myc Downstream-Regulated Gene 4 Protein), *BMP3* (Bone Morphogenetic Protein 3), β-actin, and a hemoglobin immunoassay was used to profile stool samples of about 10,000 participants with average CRC risks. The assay's sensitivities to detect CRC and advanced precancerous lesions were reported to be 92.3 and 42.4%, respectively (Imperiale et al., 2014).

In 2014, the United States Food and Drug Administration (FDA) approved a DNA-based CRC test, Cologuard (Exact Sciences Corporation, Madison, WI, United States) which is now commercially available and can detect both altered DNA and blood in the stool. Cologuard tracks aberrant methylated *BMP3* and *NDRG4* (a mutant form of *KRAS)*, beta-actin, and hemoglobin (Narayanan et al., 2014).

DNA methylation biomarkers have been under study for years, and several biomarkers have been introduced as potential markers for CRC screening. In particular, methylation events in AGTR1, WNT2, and SLIT2 genes were validated in stool DNA of CRC cases with a detection sensitivity of 78% (Carmona et al., 2013). Moreover, the determination of methylated Septin 9 (mSEPT9) in plasma has shown a remarkable sensitivity and specificity for colorectal cancers but not for adenomas (Toth et al., 2014; Nian et al., 2017). Recently, DNA methylation-based biomarkers entered into the clinical practice of noninvasive CRC diagnostics. Subsequently, a mSEPT9 assay (also referred as the Epi proColon®) received FDA approval for CRC screening in April 2016. The assay for mSEPT9 relies on qualitative detection of the methylated cytosines in Septin 9 gene-associated CpG island by Real-Time PCR in patients' blood (Issa & Noureddine, 2017).

2.2 Gut microbiome

A majority of past studies of the interaction between microbes and humans concentrated mainly on single pathogens based on Koch's postulates. Recently, global changes in the composition of microbiota (dysbiosis) were associated with the presence of certain human conditions (Schwabe et al., 2013;Turnbaugh et al., 2006; Smith et al., 2013). The microbiome is being described as the "forgotten organ" of the human body (O'Hara et al., 2006). The microbiota is estimated to comprise about one kilogram of the average adult's body weight (Savage, 1977), with about 99% of the microbiota inhabiting the gastrointestinal (GI) tract (Schwabe et al., 2013). Additionally, the human body contains about 3.72×10^{13} cells (Bianconi et al., 2013) while there are some 10^{14} bacteria cells in the human microbiome (Savage, 1977). Thus, the number of bacterial cells in the gut is

almost 10-fold greater than the number of our own cells. The gut microbiota genomes carry about 4 million genes (Qin et al., 2010), which is about 150-fold more than that in the human genome (Wu et al., 2013).

There are two bacterial communities in the colon, one is located in the lumen, while the other resides in the epithelial and cryptal regions as a mucosal biofilm that is resistant to the hydrodynamic shear forces of the colon. It has been hypothesized that dysbiosis of the mucosal communities of colon contribute to the development of inflammatory bowel disease and CRC as they are more stable than luminal communities (Savage, 1977, Sonnenburg et al., 2004) and are in direct communication with the immune system.

In 2007, The National Institutes of Health (NIH) formed the Human Microbiome Project (HMP) to characterize microbial organisms at human body sites known to have commensal bacteria and to provide reference datasets of bacterial relative abundancies and bacterial genome sequences in an attempt to define their mutual relationship in health and disease (Turnbaugh et al., 2007).

The dietary composition has an essential effect on the gut microbiota. The changes in the fecal microbiota are detectable as early as a few days after changing diets. One study showed that specific bacterial groups increased rapidly after a dietary change, but then these changes may quickly reverse upon return to previous diet (Walker, 2011). Even the short-term consumption of animal or plant diets changes microbial community structure, thus, pointing at the rapid response of the gut microbiome to dietary changes (David et al., 2014).

Different regions of the colon have dissimilar microbiota as transit time, pH, nutrient availability, oxygen exposure, host secretions, such as bile and digestive enzymes, mucosal surfaces, and interactions with the immune system vary in different parts of GI tract (Flint, 2012).

2.3 The microbiome and human health

After a decade of research in the field of the human microbiome, there are thousands of human-associated microbial strains identified allowing detailed investigations into the contribution of microbes to human health. Since 1995, associations between bacterial communities and human diseases have been studied using *in vitro* culture-based methods. Currently, these studies utilize more efficient NextGen sequencing approaches. Conventional culture-based methods detect about 30% of bacterial species (Fraher et al., 2012) while next-generation sequencing methods allow scientists to overcome this limitation.

Metabiomic research (the systems biology of the human ecosystem) is now underway to increase our understanding of the relationship between the microbiome and disease. Microbiome studies provide information about the microbiome changes related to diseases (whether it is the cause or the effect). Particularly, one can monitor the dynamics of the microbiome's diversity, relative abundance, community structure, and changes in metabolic pathways, at the metabolite, protein, and RNA levels (Morgan et al., 2014).

The microbiome is in a state of homeostasis with the human body under normal physiology and aids the host metabolic pathways, especially with respect to carbohydrate digestion, vitamins K and B production, and modulation of immunity (Dulal et al., 2014).

As such, it is now well established that microbial communities play significant roles in human health and development (De Filippo et al., 2010; Dethlefsen & Relman, 2011; Spencer et al., 2011; Dominguez-Bello et al., 2010; Koenig et al., 2011). However, imbalances of bacterial communities, known as dysbiosis, have been related to many chronic conditions, including obesity, type 2 diabetes, and autoimmune diseases such as rheumatoid arthritis, allergy, autism, and inflammatory bowel disease in addition to adenomas and CRC (Turnbaugh et al., 2009; Le Chatelier et al., 2013; Ley et al., 2006; Qin et al., 2012; Vaahtovuo et al., 2008; Russell et al., 2012; De Angelis et al., 2013; Kang et al., 2013; Wang et al., 2013; Collins et al., 2014; Kostic et al., 2014; Hold et al., 2014; Manichanh et al., 2012; Shen et al., 2010; Sobhani et al., 2011; Marchesi et al., 2011; Castellarin et al., 2012; Chen et al., 2012; Kostic et al., 2012; Sanapareddy et al., 2012; Wu et al., 2013; Geng et al., 2013; McCoy et al., 2013; Ahn et al., 2013; Brim et al., 2013; Ohigashi et al., 2013; Ohigashi et al., 2013; Weir et al., 2013; Kostic et al., 2013; Zackular et al., 2014; Zeller et al., 2014; Burns et al., 2015; Mira-Pascual et al., 2015). At least some of these associations are causal as associated phenotypes transfer with fecal transplantation experiments in germ-free mice (Carvalho et al., 2012; Turnbaugh et al., 2009, Zackular et al., 2013; Wong et al., 2017). Some studies suggest that the changes in gut microbiome could induce phenotype changes in mice (Heijtz et al., 2011; Koren et al., 2012; Smith et al., 2013). Through modulation of inflammation, apoptosis, and DNA damage, the microbiota is also involved in the etiology of many cancers (Louis et al., 2014).

2.3.1 Gut microbiota, adenoma, and colorectal cancer

Microbiome analysis is a potential screening method for CRC (Zackular et al., 2014; Eklof et al., 2017). As mentioned above, the densest and most metabolically active microbial community in healthy adults is localized in the colon. The bacterial density in the colon is about 10^{12} cells/ml while the density in the small intestine is about 10^{2} cells/ml, and correspondingly, the cancer risk in the colon is almost 12 times greater than that of the small intestine (Proctor et al., 2011; Jemal et al., 2009). Many studies have reported gut microbiome dysbiosis as a factor in the etiology of adenoma and CRC. Some of these studies are summarized in Table 1.

Table 1 A summary of CRC and adenoma microbiome studies since 2010.
There are many structural differences among these studies including sample type (tissue, stool, and rectal swab), population (various genetic background, different geography), and technical differences (Sequencing method, 16S rRNA primer); these differences can be one of the reasons that the results of these reports are not readily comparable. N/A: not available.

Study	Type of samples	disease	Population	Sample size	Method	Variable region
Shen et al. 2010	Mucosal biopsy	Adenoma	USA	21 adenoma, 23 control	Terminal restriction fragment length polymorphism, clone sequencing and fluorescent in-situ hybridization analysis of the 16S rRNA genes	N/A
Marchesi et al. 2011	Tumor/ adjacent normal tissue	CRC	Netherlands	6 CRC	Roche 454 GS FLX pyrosequencing	V1-V3
Castellarin et al. 2012	Tumor/ adjacent normal tissue	CRC	Canada	11 CRC	Illumina GAIIx, RNA seq	N/A
Chen et al. 2013	Intestinal lumen, mucosa (rectal swabs), fecal samples, tumor/ matching normal tissue	CRC	China	46 CRC, 56 control	Roche 454 GS FLX pyrosequencing	V1-V3
Kostic et al. 2012	Tumor/ adjacent normal tissue	CRC	USA & Vietnam	95 CRC	Roche 454 GS FLX pyrosequencing	V3-V5
Sanapareddy et al. 2012	Rectal mucosa biopsy	Adenoma	USA	33 adenoma, 38 control	Roche 454 GS FLX pyrosequencing	V1-V2
Geng et al. 2013	Tumor/ adjacent normal tissue	CRC	China	8 CRC	Roche 454 GS FLX pyrosequencing	V1-V2
McCoy et al. 2013	Rectal mucosa biopsy	Adenoma	USA	48 CRC, 67 control	Roche 454 GS FLX pyrosequencing	V1-V3
Zeller et al. 2014	Tumor/ adjacent normal tissue	CRC	Germany	38 CRC	llumina MiSeq	V4
Burns et al. 2015	Tumor/ adjacent	CRC	USA	44 CRC	Illumina MiSeq	V5-V6

Study	Type of samples	disease	Population	Sample size	Method	Variable region
	normal tissue					
Mira-Pascual et al. 2015	Fecal and biopsy	CRC & adenoma	Spain	7 CRC, 11 tubular adenoma, 10 control	Roche 454 GS FLX pyrosequencing	N/A
Nakatsu et al. 2015	Biopsy	CRC & adenoma	China	52 Tumor/adjacent, 47 adenoma/ adjacent, 61 control	Roche 454 GS FLX pyrosequencing	V1-V4
Thomas et al. 2016	Biopsy	CRC	Brazil	18 rectal cancer, 18 control	Ion-torrent PGM platform	V4-V5
Xu & Jiang 2017	Biopsy	CRC & adenoma	China	52 cancer, 47 adenoma, 61 control	Roche 454 GS FLX pyrosequencing	V1-V4
Gao et al. 2017	Tumor/ adjacent normal tissue	CRC	China	65 CRC	Illumina MiSeq	V4
Yoon et al. 2017	Biopsy	CRC & adenoma	Korea	6 CRC, 6 conventional adenoma, 6 sessile serrated adenoma, 6 control	Roche 454 GS FLX pyrosequencing	V1-V3
Hale et al. 2017	Fecal	Adenoma	USA	233 adenoma, 547 control	Illumina MiSeq	N/A

Bacterial dysbiosis, the state of increased abundance of unfavorable (allochthonous) and decreased abundance of beneficial (autochthonous) bacteria, has been associated with adenoma and CRC in all of these studies (Keku, 2015). Moreover, it has been revealed that in CRC patients, the compositions of the microbial community of the gut are stage-specific (Kinross et al., 2017).

No specific bacterial species can be used as a universal biomarker for CRC. One possible reason for the lack of specific microbiome biomarkers is the complexity and dynamics of the system which is heavily influenced by diet, inflammation, host genetics, and the unique microbiome structure in each individual. However, an uncommon phylum,

Fusobacterium, particularly, the *Fusobacterium nucleatum*, has been detected more often in CRC patients than in controls (Zackular et al., 2014; Kinross et al., 2017). It is suggested that the abundance of *F. nucleatum* may be helpful in predicting clinical outcomes in patients with CRC. A significant association between the abundance of *F. nucleatum* and the size and stage of CRC was reported. Moreover, the survival of patients with CRC correlates negatively to the abundance of this species (Yamaoka et al., 2017).

The hope is that with improved data analysis approaches that examine all the disease factors, we can distinguish healthy, adenoma, and CRC associated microbial communities.

2.4 16S rRNA analysis for microbial studies

Currently, Small Subunit rRNA (SSU rRNA) gene sequencing is performed to detect and compare bacterial populations in environmental and human samples including blood, serum, tissue, and stool. The SSU rRNA is alternatively known as 16S rRNA in bacteria. The approach allows one to identify the structure and diversity of the microbiome by determining the abundance, phylogeny, and taxonomy of samples from complex microbiomes in environmental samples.

16S rRNA sequences are essential in studies of microbial ecology and evolution. In 1987, Carl Woese founded the field of molecular evolution when he started sequencing the small subunit rRNA genes (SSU rRNA) to determine evolutionary distances. Using this method, he developed the revised tree of life with a newly identified domain named "Archaea" (Woese, 1987). Currently, the 16S rRNA gene is the standard used to determine phylogenetic relationships for bacteria, to assess diversity in the environment, and to detect

and quantify specific bacterial populations (Acinas et al., 2004). The original reason for choosing 16S rRNA was that it was the easiest nucleic acid molecule to purify. One merely spun down the ribosomes and extracted the RNA. Fortuitously, this molecule has a number of unique features that make it ideal for evolutionary studies. The length of the 16S rRNA gene is 1550bp (Clarridge, 2004) which is a suitable length for diversity analysis as it has been suggested that 500-700 bp are sufficient to allow sequence assignments at the species level (Clarridge, 2004). The conserved and variable regions in the 16S rRNA sequences facilitate the computational aspects of studying diversity and evolution (Van de peer et al., 1996; Jonasson et al., 2002; Acinas et al., 2004). The nine variable loop regions are located between the conserved helical regions and the conserved helical regions are good targets for primer design while the variant loop regions, V1-V9, provide the signal to differentiate species diversity (Lane et al., 1985). Additionally, the 16S rRNA gene is present in all prokaryotes on the planet and has the same function in all cells. It carries both fast- and slow-evolving regions that give us the opportunity to find closely and distantly related species and strains. The helical regions are sufficiently conserved to allow accurate alignments among all species and the variable loops allow evolutionary and diversity studies. Finally, 16S rRNA genes are good targets for evolutionary studies because horizontal gene transfer rarely or never affects these genes (Acinas et al., 2004). In addition, many reference databases are available for 16S rRNA sequences that make taxonomy studies feasible. The main limitation of 16S rRNA sequencing is that this molecular approach will detect both live and dead cells and there is not a direct correspondence between taxonomic identity and functionality.

Despite these limitations of the16S rRNA analysis, it is the most popular method for microbiome analysis because it is convenient, fast, and cost-effective (Preheim et al., 2013). As noted above, the rRNA gene sequencing is the most common technique used for most studies in the Human Microbiome Project (www.hmpdacc.org), the Earth Microbiome Project (www.earthmicrobiome.org), and many other individual projects worldwide.

The most challenging and time-consuming step in the 16S rRNA analysis starts after sequencing, when bioinformatics and statistical analyses are used to extract pertinent information from large numbers of sequences or reads (Ju and Zhang, 2015). The processing of 16S rRNA gene sequences in most software pipeline can be divided into three steps: preprocessing, selection of Operational Taxonomic Units (OTUs), statistical analysis and visualization. The preprocessing step includes de-multiplexing which is the process of assigning reads to the corresponding barcoded samples and removing the sample barcodes (Gillevet et al., 2008), quality filtering, denoising, chimera removal, and data normalization. After preprocessing, the resulting "clean" or "effective" reads are then used for selecting representative sequences that define OTUs. At the next step, representative sequences are aligned to reference databases for taxonomic assignment. A phylogenetic tree of the OTUs is made at the end of this step. The final step is advanced data analysis and visualization, including calculations of alpha and beta diversity, ordination analysis, clustering and classification, and data visualization which includes heatmaps, principal coordinates analysis plots, and networks (Ju and Zhang, 2015). These steps of 16S rRNA analysis will be discussed in details in following sections.

2.4.1 Sequencing 16S rRNA genes

The first step in a 16S rRNA survey is to construct a library of DNA fragments and sequence them. Several sequencing systems with different technologies are available for sequencing studies including the Roche 454 GS FLX (based on pyrosequencing technique), Illumina (based on sequencing by synthesis), Applied Biosystems SOLiD (sequencing by ligation), Pacific Biosciences (single-molecule real-time (SMRT) sequencing by synthesis), Ion Torrent (semi-conductor sequencing), Helico (single-molecule sequencing by synthesis), and Oxford Nanopore (single strand sequencing). These methods differ based on their technology, runtime, read length, sequencing yield, coverage, accuracy, and cost (Kuczynski et al., 2012; Rees et al., 2012; Oulas et al., 2012; Liu et al., 2013; Rizzo et al., 2012; Weirather et al., 2017). The 454 technology is no longer supported by the manufacturer. However, there are still many datasets produced by this technology that are available online and used by researchers in meta-analyses. A summary of the characteristics, advantages, and disadvantages of these systems is provided in Table **2**.

Table 2 A summary of 16S rRNA sequencing technologies, their characteristics (Table 2A), advantages, and disadvantages (Table2b).
These technologies differ in many aspects such as accuracy, read length, and duration of runs.

Table 2A Sequencing platform	Technology	Year released	Paired -end	Read length	Accuracy (single read not consensus)	Read/ run	Sequence yield per run	Time per run
Sanger sequencing	PCR; dideoxy chain termination	1997; automated version: 2003	No	400 to 900 bp	99.999%	96	1.9-84kb	3 h
454 GS FLX	Emulsion PCR; pyrosequencing	454 (2005-2014)	Yes	700-1000 bp	99.90%	10^6	0.7G	10-24 h
Illumina_Solexa		2007				3x10^6		
SOLiD sequencing	Emulsion PCR; ligation and two-base coding	2008		50+35 or 50+50 bp	99.9% (99.94%)	10^8	120G	7-14 days
Ion Torrent sequencing	Emulsion PCR; ion semiconductor	2010	Yes	up to 400 bp	98%	1.5-3x10^6	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	2-4 h
Illumina-MiSeq	Bridge PCR; sequencing by synthesis	2011	Yes	300 bp x 2	99.9% (Phred30)	25x10^6	1.5-15G	1 to 11 days, depend upon sequence & read length
Illumina-HiSeq	Bridge PCR; sequencing by synthesis(reverse terminator)	2012	Yes	125bp x2	99.9%	3x10^9	600-1000G	3(single end)-10 (pair end) days
Helico	No amplificatio	2012		55bp	~100% accuracy	21-35G		8d

Table 2A Sequencing platform	Technology	Year released	Paired -end	Read length	Accuracy (single read not consensus)	Read/ run	Sequence yield per run	Time per run
	n; single-molecule sequencing (sequencing by synthesis)							
Pacific Biosciences	No amplification; single-molecule real-time sequencing (sequencing by synthesis	2012		10,000 bp to 15,000 bp avg (14,000 bpN50); maximum read length> 40,000 bases	87% single-read accuracy		50,000 per SMRT cell, or 500–1000 megabases	30min - 4h
Oxford Nanopore (GridIon)	No amplification; label-free single molecule real-time sequencing	2012		10000bp		100G		5h

Table 2B		
Sequencing platform	**Advantages**	**Disadvantages**
Sanger sequencing	Long individual reads, high quality, useful for many applications	Low throughput, more expensive and impractical for larger sequencing projects, this method also requires the time-consuming step of plasmid cloning or PCR
454 GS FLX	Long read size, fast run	Expensive runs, high reagent cost, homopolymer errors, low throughput, prone to base insertion and deletion errors during base calling but rare chance of substitution errors, no longer supported
SOLiD sequencing	Low cost per base, accuracy	Slower than other methods (long run), has issues in sequencing palindromic sequences, short read assembly
Ion Torrent sequencing	Less expensive equipment, fast	Homopolymer errors, higher quality than 454 especially when sequencing homopolymers
Illumina-Miseq	Potential for high sequence yield, depending upon sequencer model and desired application	Expensive equipment, requires high concentrations of DNA, higher substitution error rates, lagging strand dephasing causes sequence quality deterioration towards the end of read
Illumina-Hiseq	High throughput	Short read assembly, long run time, all samples on flow cell sequenced at same read length, higher substitution error rates, lagging strand dephasing causes sequence quality deterioration towards the end of read
Helico	No amplification bias, shorter preparation time, expensive	Machine not widely used; sequencing service available through company
Pacific Biosciences	Long read length, fast, detects 4mC, 5mC, 6mA	Moderate throughput, equipment can be costly, high error rate
Oxford Nanopore (GridIon)	Potentially sequence entirely intact DNA strands/polymers, eliminate erroneous sequencing caused by shotgun metagenomics and exclude the need for the error-prone assembly step during data analysis	Low throughput; high error rate

Two of the most popular systems in the last few years were the Roche 454 GS-FLX and the Illumina (Oulas et al., 2015). The 454 pyrosequencing utilizes emulsion PCR where DNA library fragments are immobilized on beads, then amplified in oil droplets. The clonally amplified beads are then spun into wells on a picotiter plate and sequencing by synthesis is performed. Specifically, the four nucleotides are added sequentially and iteratively in a cyclic manner and pyrophosphate (PPi) is released after incorporation of each nucleotide. The released PPi reacts with adenosine 5´phosphosulfate (APS) in the presence of ATP sulfurylase to produce ATP. The ATP is then utilized by luciferase-mediated conversion of luciferin to oxyluciferin to produce light that is detected by a charged-coupled device (CCD) camera. The intensity of the light is proportional to the number of nucleotides incorporated (up to detector saturation) (Mardis, 2008; Ronaghi, 2001). Thus, the pyrosequencing method relied on the released pyrophosphates to detect an incorporation event, in contrast, the fluorescent chain-terminating dideoxynucleotides utilized in the Sanger method. This was one of the first commercially implemented NextGen technologies. However, the Roche pyrosequencing machines are no longer supported by the manufacturer as the bead size is relatively large. This technology has been replaced by the emulsion-based Ion Torrent technology where incorporation events are detected by the proton released using a semi-conductor chip.

The Illumina technology uses a sequencing-by-synthesis approach as well. However, at the beginning of the process, DNA is purified and then tagmented (tagging of the double-stranded DNA with a universal overhang) by transposons which randomly cut DNA into short pieces. Sequencing adapters and sample barcodes are then added on either

side of the DNA fragments by PCR. Sequencing begins with pairing DNA molecules to complementary oligonucleotide adapters attached to a slide. Bridge amplification of fragments occurs on this oligo-derivatized surface via DNA polymerase producing multiple DNA copies that form clusters or polonies (polymerase colonies). This process occurs in a machine called a "cluster station." Each cluster contains about one million copies of the original DNA fragment. Sequencing by synthesis is performed using four nucleotides that are fluorescently labeled and bound to a blocking group. They are added simultaneously to the flow cell channels and, after each nucleotide is incorporated, a laser excites the dyes and the CCD camera photographs the incorporation event. Then the incorporated base in each cluster is identified after color deconvolution. Subsequently, the 3' terminal blocks are removed and the next nucleotide is added to the sequence. The process continues until all the fragments are sequenced (Oulas et al., 2015).

Illumina offers a variety of sequencing instruments for different applications including MiSeq and HiSeq. The Miseq instrument was developed for longer reads (about 300 bp) but the output is lower (25 million paired-end sequencing reads of 300 bp in length or 15 GB) than other instruments. The HiSeq platform is suitable for sequencing up to 125 bp but its output is much higher (1,000 GB per run). The reagent cost of Illumina sequencing is less than the 454 pyrosequencing but the runtime is longer. The sample preparation size is 20 ng of DNA for both pyrosequencing and Illumina technology. The shorter reads in Illumina (300 bp compared to 700 bp in pyrosequencing) may increase the time of analysis as more noise removal algorithms are necessary. On the other hand,

Illumina has lower sequencing error rates reportedly. Currently, Illumina is more popular than the ion torrent technology (Oulas et al., 2015).

2.4.2 Preprocessing the sequences

After obtaining the sequencing reads, preprocessing of the reads is performed. Methodological artifacts such as polymerase and sequencing errors, chimeras, and primers and barcodes must be removed from the sequences in order to have high quality reads for OTU selection. All raw barcoded sequences are demultiplexed, labeled with samples identifiers, and then removed. The next step of this preprocessing is quality filtering where sequences with poor quality bases or mismatches are removed to prevent the confounding of downstream analyses. Factors that are considered for quality filtering are minimum average quality scores, the maximum number of ambiguous bases, minimum and maximum sequence lengths, maximum lengths of homopolymers, and maximum mismatches in primers or barcodes (Ju and Zhang, 2015). Noise in the sequence data due to errors that occur during PCR amplification such as sequencing errors, PCR single base substitutions, and PCR chimeras should be removed as these can inflate estimates of alpha diversity in microbial communities (Reeder and Knight, 2010; Quince et al., 2011). Chimeras are hybrid PCR products that result from copying multiple parent sequences that can be falsely interpreted as novel organisms (Haas et al., 2011). Software tools are available for noise and chimera removal such as Denoiser (in QIIME), AmpliconNoise (including PyroNoise and SeqNoise), Acacia, Pre.cluster (in Mothur), ChimeraSlayer, UCHIME, Perseus, and DECIPHER (Quince et al., 2011; Bragg et al., 2012; Haas et al., 2011; Edgar et al., 2011; Quince et al., 2011; Wright et al., 2012). These methods differ on

the algorithm they use to find the chimeras, speed, and sensitivity. In this project, UCHIME was used for chimera detection. UCHIME algorithm looks for a 3-way alignment of a query sequence with two parent sequences where one of the parent sequences is more similar to one region of the query, and the other parent is similar to another region as illustrated in Figure 2. A score is calculated for each alignment and a higher score indicates a stronger chimeric signal (Edgar et al., 2011).

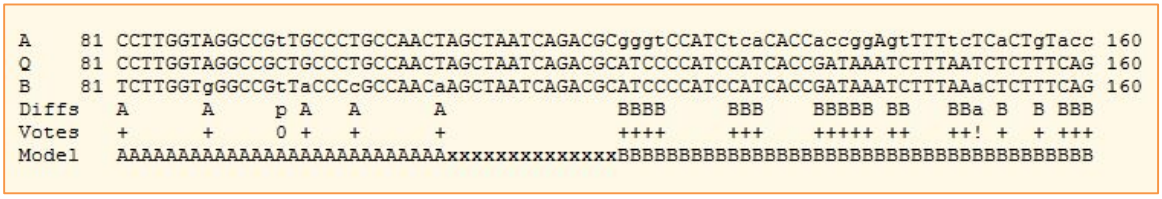


Figure 2 A region from a chimeric alignment by UCHIME.
A 3-way alignment of a query sequence with two parent sequences. "Diffs" show the different bases between query and parents. Votes are yes (+), no(!), and abstain (0). The "Model" shows the final pattern of chimera detected. Ref: Edgar et al. (2011).

Sequencing depth, i.e., the number of clean (preprocessed) sequences obtained for each sample, can vary among different barcoded samples due to initial sample pooling inconsistencies. As such, different numbers of reads for different samples will affect the diversity estimation in downstream analyses. Therefore, data normalization is recommended. Various methods are available to normalize the reads including rarefaction, relative abundance, and Z-score (Ju and Zhang, 2015). Rarefaction is the random selection of an equal number of sequences from each sample, and this number is chosen as the minimum sequence count for all samples. Relative abundance is read counts of a taxon against the total sample read counts (Goodrich et al., 2014). In the Z-score method, a score

is calculated as the difference between the observed and mean values divided by the standard deviation (Oswald et al., 2011).

2.4.3 Creating Operational Taxonomic Units (OTUs)

The next step in the analytical process is to generate Operational Taxonomic Units (OTUs). We define an OTU as a distinct taxonomic entity. However, it may not have a known taxonomic name. One method of OTU identification is the "*de novo* OTU selection" method in QIIME. In this approach, OTUs are created based on the similarity of reads to each other. Another method, known as the "close-reference method," is to align the reads to a 16S rRNA reference database such as "GreenGenes"(http://greengenes.lbl.gov) and designate OTUs based on the similarities of reads to known taxa. This latter approach is quick and convenient. However, novel OTUs may be disregarded. An alternative is the "open-reference method," in which the *de novo* and close-reference methods are combined (Preheim et al., 2013). The benefit of the *de novo* OTU selection method is that all reads are clustered and no sequences are lost. A drawback is the computational speed as it can be too slow to apply to large datasets (e.g., more than 10 million reads). However, if a reference database is not available for the desired sequences, this method should be used. This method is not applicable when comparing non-overlapping amplicons such as the V2 and V4 regions of the 16S rRNA and does not recommend for the processing of the large datasets.

As mentioned above, the closed-reference approach is much faster than the *de novo* algorithm, and the trees and taxonomy results of this approach will be more accurate. In addition, since all OTUs are produced from a standard reference, closed-reference data

may be compared with the results generated in other studies. The disadvantage of the closed-reference method is the loss of novel diversity as all non-matched reads are discarded. An open-reference method is preferable to *de-novo* and closed-reference because it has the advantages of both methods (Rideout et al., 2014) as the open-reference OTU selection algorithm runs de novo clustering on the sequences that failed to match with the reference database and adds them to the analysis. This approach decreases the runtime and the open-reference OTU selection method can be applicable to billions of reads (Rideout et al., 2014).

There are many *de novo* approaches available. In this study, we used the UPARSE, UPGMA, and UCLUST algorithms. These methods are the default methods of the most commonly used 16S rRNA sequencing pipelines such as QIIME (UCLUST), Mothur (UPGMA) and USEARCH (UPARSE) (Caporaso et al., 2010; Schloss et al., 2009; Edgar, 2013). The aim of using these three methods was to compare them for their accuracy in OTU selection and determine which pipeline was the most effective for classification of subjects.

UPARSE: The UPARSE pipeline was developed by Edgar (Edgar, 2013) and uses a "*de novo"* clustering method that is claimed to work faster and more accurate than other commonly used OTU clustering methods for microbial studies. It starts with quality-filtering the reads and trimming them to a fixed length. The singletons can be optionally discarded, and the remaining reads are used for clustering. The clustering algorithm is called "UPARSE-OTU" which is a 'greedy' algorithm that performs simultaneous chimera filtering and OTU clustering. UPARSE is highly robust with respect to variations in the

input data order and can be successfully applied to a wide range of marker genes and sequencing technologies.

The goal of UPARSE-OTU is to identify a set of representative sequences (OTUs) satisfying the following criteria as shown in Figure 3.

- All pairs of sequences in an OTU should have at least 97% pair-wise sequence identity.
- Chimeric sequences should be discarded.
- All non-chimeric input sequences should match at least one OTU with ≥ 97% identity (Edgar, 2013).

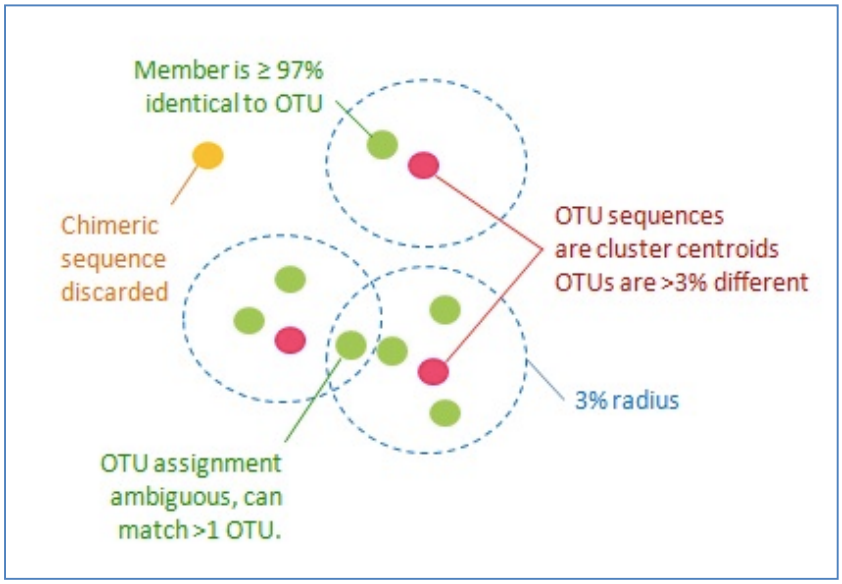


Figure 3 UPARSE-OTU clustering criteria.
All pairs of OTU sequences have at least 97% pair-wise sequence identity. Chimeric sequences are discarded.
All non-chimeric input sequences match at least one OTU with ≥ 97% identity.
Ref: http://drive5.com/usearch/manual/uparseotu_algo.html

UPGMA (Unweighted Pair Group Method with Arithmetic Mean): The UPGMA method is an agglomerative "bottom-up" clustering method that creates hierarchical clusters. It begins with the creation of one cluster for each of the input reads. Then the closest two clusters are identified and joined into a higher-level cluster as showed in Figure 4. In the UPGMA method, the average linkage will be calculated as the distance between clusters when one of those clusters has more than one sequence.

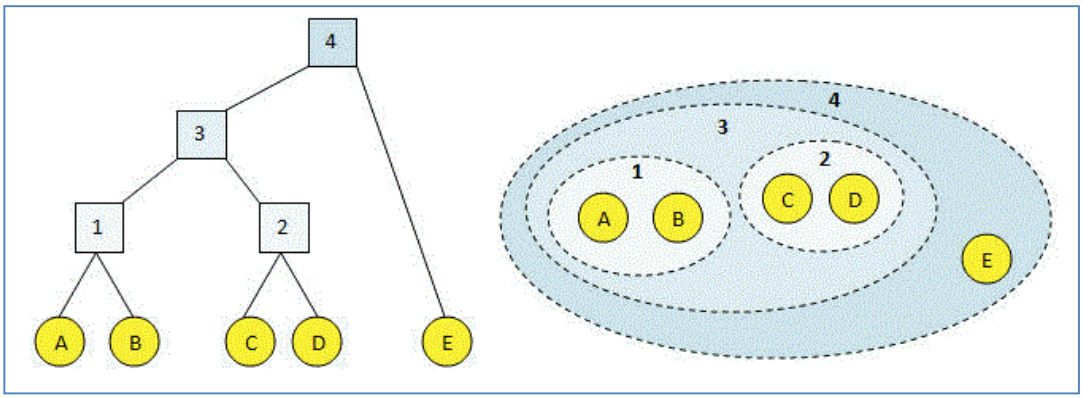


Figure 4 UPGMA algorithm.
This algorithm is a bottom-up clustering method that creates one cluster for each of the input reads. Then the closest two clusters joined. This process continues until one single cluster remains. Ref: http://drive5.com/usearch/manual/agg.html

UCLUST: The UCLUST method clusters the sequences into OTUs based on their identities. Each cluster has a centroid or representative sequence and the sequences in a cluster have similarities greater than or equal to a defined identity threshold while sequences out of the cluster have similarities less than that identity threshold as depicted in Figure 5. Thus, centroids have similarities to each other that are less than the threshold. However, the order of the input sequences is critical in UCLUST as the sequences are put

into the clusters in the order in which they appear in the sequence file. Specifically, the first sequence is the first centroid. If the next sequence has an identity greater than the threshold of the first one, it will be placed in the same centroid. Otherwise, it becomes a new centroid.

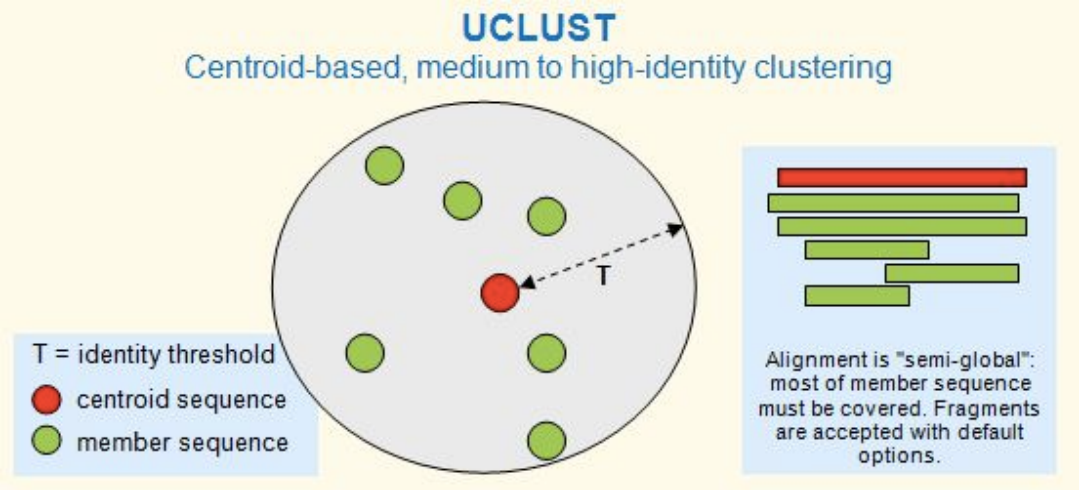


Figure 5 UCLUST algorithm.
UCLUST make clusters based on the identity of sequences to each other. Ref: http://drive5.com/usearch/manual/uclust_algo.html

USEARCH (Edgar, 2010): An additional function of UCLUST is to find the similarities of input sequences with the centroids identified so far. This function is performed by the USEARCH algorithm where identity is defined as the number of identities in each alignment column divided by the total number of alignment columns which is similar to the identity score used in the BLAST (Edgar, 2010). The USEARCH algorithm searches sequences for high-identity hits to one or more sequences ("targets") as

shown in Figure 6. USEARCH is effective in detecting identities greater than about 75% for nucleotides. For lower-identity alignments, the local search UBLAST is used.

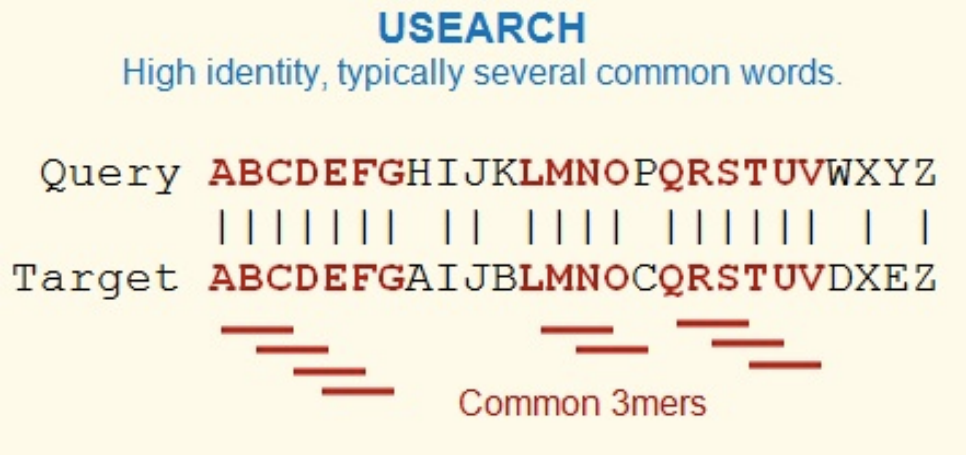


Figure 6 USEARCH algorithm.
USEARCH exploits the fact that similar sequences tend to have several short words in common, and uses the word count to prioritize the database search.
Ref: http://www.drive5.com/usearch/manual/usearch_algo.html

Choosing the OTU selection method depends on the dataset and the aim of the study.

2.4.4 Assigning taxonomy to OTUs

Labeling the OTUs with a taxonomic identity is the next step of the analysis pipeline. The Ribosomal Database Project (RDP), GreenGenes (DeSantis et al., 2006), SILVA, (Quest et al., 2013), and GEBA are the references databases that can be used for assigning taxonomy to the OTUs.

The Ribosomal Database Project (RDP) (https://rdp.cme.msu.edu) provides quality-controlled, aligned, and annotated bacterial and archaeal 16S rRNA sequences as

well as fungal ITS and 28S rRNA sequences for Bayesian Classification, annotation, and alignment (Cole et al., 2014).

GreenGenes (http://greengenes.lbl.gov) is a bacterial and archaeal 16S rRNA sequence database that provides annotated, chimera-checked, full-length 16S rRNA gene sequences in standard alignment formats which allows researchers to use this database as a reference for the taxonomic identification of OTUs and distinguish chimeras (DeSantis et al., 2006).

SILVA (http://www.arb-silva.de) is a comprehensive online resource that consists of fully aligned and regularly updated small (16S/18S, SSU) and large (23S/28S, LSU) subunit rRNAs of Bacteria, Archaea, and Eukarya. SILVA can be used to check the quality of the reads and for the taxonomic identification of OTUs (Pruesse et al., 2007; Quast et al., 2013).

The Genome Encyclopedia of Bacteria and Archaea (GEBA) (http://www.jgi.doe.gov/programs/GEBA/) is mostly used to fill in the taxonomic gaps for bacterial and archaeal databases. These taxonomic gaps emerge because of the highly biased phylogenetic distribution in the available genome sequences when compared to the extent of the total environmental microbial diversity known today. The GEBA group is trying to provide genomes for all bacterial and archaeal groups (Wu et al., 2009).

2.4.5 Finding Phylogenetic relationships among OTUs

To visualize the phylogenetic relationships between OTU sequences, they are first preprocessed and then trimmed to the same length which is defined as the most frequent length of the population. Duplicates are removed, and the trimmed sequences are aligned

to each other. In this study, mafft (Katoh et al., 2002) was used for sequence alignment as the CPU time is reduced dramatically with this algorithm while it has comparable accuracy to other methods (Katoh et al., 2002). The next step is producing a phylogenetic tree. A phylogenetic tree is a branching diagram showing the inferred evolutionary relationships among various OTUs based on their sequence similarities. Thus, the taxa that are descended from a common ancestor will be connected in the cladogram (Ju and Zhang, 2015). In the present study, FastTree (Price et al., 2009; Price et al., 2010) was used for constructing the phylogenetic trees. There are two versions of FastTree, FastTree1 and FastTree2. FastTree 1 employs nearest-neighbor interchanges (NNIs) and the minimum-evolution criterion to improve the tree while FastTree 2 adds minimum-evolution subtree-pruning-regrafting (SPRs) and maximum likelihood NNIs. FastTree 2 uses heuristics to restrict the search for better trees and estimates a rate of evolution for each iteration. For large alignments, FastTree 2 is 100–1,000 times faster than FastTree1. The phylogenetic tree produced at this step is used for downstream beta diversity analyses such as UniFrac (Lozupone et al., 2005) as described below.

2.4.6 Alpha and beta diversity analyses

Alpha diversity: Estimating the alpha diversity (α diversity), which is the diversity within samples, is the first step of community analysis as it is crucial in describing the structure, function, and evolutionary patterns within communities. To determine the α diversity, species richness (i.e., the number of species) and the relative abundancies of the different species are measured (Ju and Zhang, 2015).

The accuracy of diversity measurement is dictated by the depth of sequencing (i.e., the number of reads per sample). The sequencing depth dictates the diversity within a sample as the higher number of reads in the sample will detect higher diversity. Because sequencing depth varies between barcoded samples due to sample pooling artifacts, rarefaction or other standardization methods should be applied before samples are compared to each other. Thus, rarefaction analysis is necessary to capture the total diversity within the sample.

Many metrics are available to estimate community diversity, such as Shannon, Simpson, invSimpson, and sobs indices (Hill et al., 1973), all of which measure both richness and evenness. The number of species per sample is a measure of richness and, thus, the more species present in a sample, the 'richer' the sample. Evenness is a measure of the normalized relative abundance of the different species. The Simpson's diversity index can be thought of as the probability that upon randomly choosing an OTU from a sample, the OTU has already been observed. Therefore, higher Simpson's diversity index indicates less diversity. The invSimpson index is the inverse of the classical Simpson's diversity index. The Simpson-based metrics are not affected by sampling effort while the Shannon index is. Observed species or observed OTUs metric (sobs) is a simple metric that just counts the number of OTUs that are present in the given sample, as abundance is not considered in this metric. Samples with the same sobs value have a similar richness and samples with higher sobs value have higher richness.

Beta diversity: Beta diversity is the diversity between communities. β-diversity metrics can be categorized into two types; qualitative where only the species diversity is

used (e.g., unweighted-UniFrac) (Goodrich et al. 2014) and quantitative where abundance is also included in the diversity measurement (e.g., Bray-Curtis and weighted-UniFrac metrics).

UniFrac: Unique Fraction (UniFrac) is a β-diversity metric that employs phylogenetic information to compare microbial communities. Accompanied with standard multivariate statistical approaches such as principal coordinates analysis (PCoA), UniFrac can describe differences between microbial communities. The UniFrac metric measures the difference between two communities based on the amount of unique evolutionary history found in the two communities. Two versions of this program exist; weighted and unweighted. In the weighted version, the relative abundancies of sequences are taken into account along with the phylogenetic similarities (shared branch length) to calculate how similar the communities are. The unweighted version only uses the tree topology to calculate the metric (Lozupone et al., 2007).

The UniFrac distance between a pair of samples is the sum of the branch length that was observed in one sample (the *unique* branch length) divided by the sum of the branch length that was observed in either sample (the *observed* branch length). The unweighted UniFrac distance between a pair of samples A and B is defined as follows:

$$\mathbf{UAB}=\frac{unique}{observed}$$

Equation 1 Unweighted UniFrac distance between a pair of samples A and B

where “unique” is the unique branch length, or branch length that only leads to OTU(s) observed in sample A or sample B, and observed is the total branch length observed in either sample A or sample B.

The UniFrac metric can determine whether the phylogenetic lineages between samples are different and can be used to cluster samples via multivariate statistical methods. UniFrac is integrated into the QIIME and Mothur microbial analyses pipelines (Ju and Zhang, 2015).

Bray-Curtis: Bray-Curtis metric (Bray & Curtis, 1957) is a quantitative non-phylogenetic based β-diversity metric that quantifies the compositional dissimilarity between two different samples based on counts at each of sample. The Bray-Curtis dissimilarity between a pair of samples, j and K, is defined as follows:

$$\mathbf{BCjk}=\frac{\sum_i|Xij-Xik|}{\sum_i|Xij+Xik|}$$

Equation 2 Bray-Curtis dissimilarity

i: feature (e.g., OTUs)

X_{ij}: frequency of feature i in sample j

X_{ik}: frequency of feature i in sample k

The calculations are performed for each pair of samples, and a dissimilarity matrix containing all pairwise distances is made. Bray-Curtis dissimilarity matrix can be used to compare community dissimilarity based on OTU abundance.

Principal component analysis (PCA) and Principal coordinates analysis (PCoA): The PCA algorithm uses an Eigen analysis to find new sets of dimensions that capture the data variability. The first dimension is chosen in a way that captures the maximum possible variance of the data while the second dimension is chosen orthogonal to the first dimension to capture as much of the remaining variance as possible. The Eigen analysis transforms potentially correlated features into fewer components known as "principal components" or "Eigen vectors." Each axis has an eigenvalue that is related to the amount of variance explained by the axis. Thus, the first Eigen vector has the highest Eigen value and this axis explains the greatest fraction of the variance of the data. The second Eigen vector has the second highest Eigen value and this axis explains the next greatest fraction of variance (Tan et al., 2006). In this way, the dimensionality of the features (e.g. OTUs) is reduced and projected along Eigen vectors representing the largest the variance of the features. To visualize the PCA results, usually, only two PCA axes are plotted as the third axis is usually less informative than the first two. The PCA plot allows rotation the cloud of data points to visually inspect the clustering.

PCoA is a similar ordination technique that employs an appropriate distance matrix. As above, it plots samples along Eigen vectors reducing dimensionality while preserving their distance relationships as much as possible. Thus, PCoA can be employed in microbiome studies to visualize similarities or dissimilarities of microbial communities using a suitable distance matrix instead of the covariance matrix used by PCA (Kindt et al., 2005).

2.5 Statistical analysis of microbiome data

The high throughput nature of next-generation sequencing allows the parallel sequencing of a large number of barcoded samples. Non-parametric statistical approaches can be used to find the association of OTUs with metadata and to determine any association between bacterial species and clinical or other variables. Many statistical tests are available to find the significant differences between two groups or classes (e.g., healthy and cancer). Some examples are the Wilcoxon signed-rank test, Kruskal–Wallis test, and the Mann-Whitney U test for non-parametric data.

MetaStats (White et al., 2009) is a statistical approach designed to identify differentially abundant features in metagenomic and 16S rRNA sequence datasets. This program utilizes the nonparametric t-test, Fisher's exact test, and the false discovery rate (FDR) to provide users with a prioritized list of remarkable features that define differences between two classes (e.g., healthy vs. ill).

The analysis of variance (ANOVA) test is used to measure significant differences between means of multiple independent samples with normally distributed data and equal variance. If the data are not normal, other nonparametric tests, such as the Kruskal-Wallis test or PerANOVA can be used (Ju and Zhang, 2015). The Kruskal–Wallis test uses a rank-ordered One-way ANOVA and is a method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends to the Mann–Whitney U test when there are more than two groups. A significant Kruskal-Wallis test indicates that at least one sample stochastically dominates over at least one other sample. The test does not identify where this stochastic dominance occurs or how many pairs of groups contribute to the stochastic

dominance. Dunn's test can be used to analyze specific sample pairs for stochastic dominance. For example, if the researcher can make the less stringent assumptions of an identically shaped and scaled distributions for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

The Linear discriminative analysis effect size (LEfSe) (Segata et al., 2010) is an algorithm for high-dimensional biomarker discovery and detection of genomic features such as genes, pathways, and taxa that can characterize the differences between two or more biological classes. This algorithm is helpful to identify differentially abundant features that are also consistent with biologically meaningful categories (classes) by taking into account both statistical significance and biological relevance. It first detects statistically different features using the non-parametric Kruskal-Wallis sum-rank test and then using another pairwise test (Wilcoxon) to determine whether the detected differences are consistent with biological behavior. To estimate the biological effect of each differentially abundant feature, they used a linear discriminant analysis (LDA).

The indicator metric is another statistically-based tool developed to find the indicator species (e.g. OTUs) that define a distinctive aspect or characteristic of an environment. Indicator combines the species relative abundance with the relative frequency of occurrence in various classes. When all the individuals of a species are represented in one group, and all the species appear in all the samples of that group, the indicator index is defined as high. A randomization method is used to find the statistical significance of the

metric. The indicator index for each species is independent of the other species relative abundance and is independent of classification approaches (Dufrene et al., 1997; McCune et al., 2002).

2.6 Data mining and machine learning for microbiome data

The development of next-generation sequencing has led to a significant decrease in the cost of sequencing (Sboner et al., 2011) and this reduction in cost has facilitated large-scale studies. However, the corresponding amount of data produced from these studies is so enormous that interpreting the data has become slow, confounded, and challenging. With the estimation of data doubling every few months, novel techniques should be developed in order to utilize the overwhelming amount of data efficiently. Data mining is a method to find previously unknown information patterns in data resources (Witten et al., 2005). Data mining extracts patterns and finds solutions to problems in large datasets. The process must be automatic or semiautomatic, should be inexpensive, and the discovered pattern should present a meaningful result (Witten et al., 2005). Data mining is a part of the knowledge discovery process that analyzes data and applies algorithms to generate patterns or models (Fayyad et al., 1996) as shown in Figure 7.

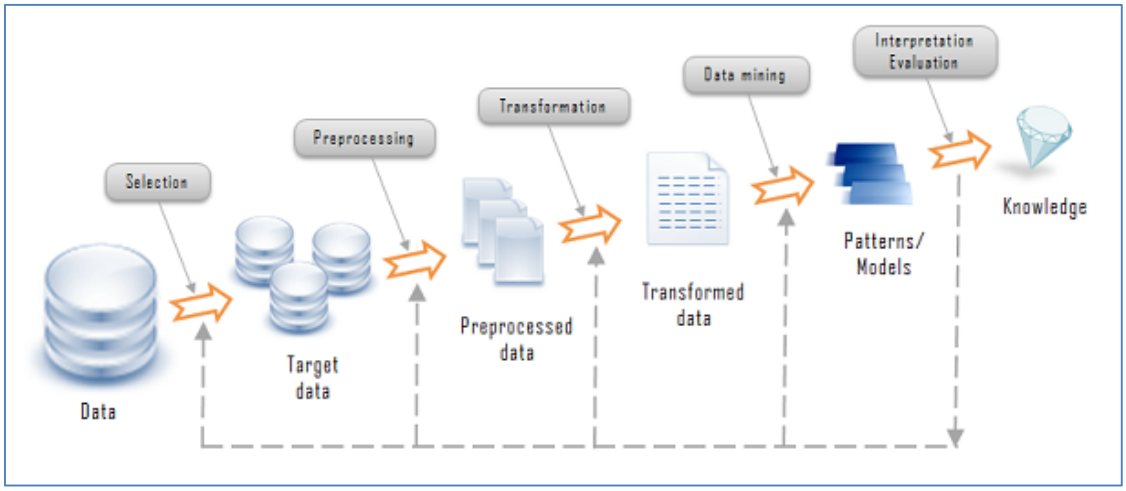


Figure 7 Data mining process.
Data mining is a part of the knowledge discovery process that analyzes data and applies algorithms to generate patterns or models. Ref: Fayyad et al., 1996.

Data mining procedures can be either "unsupervised," in which the "class" is unknown or undiscovered, or "supervised" for which the "class" is known *a priori*. Classification and regression are two examples of supervised predictive modeling tools.

Machine learning involves the study of algorithms that can extract useful information automatically. Some of these procedures may include the ideas derived from or inspired by, classical statistics.

Statistical learning methods are currently in use to identify associations in the metadata. In statistical learning, machine learning methods are applied to find a predictive pattern based on the input data. These techniques have some advantages over statistical approaches as they can detect nonlinear associations between metadata and microbiome species or OTUs.

Clustering and classification are two statistical mining methods that are widely used in taxonomic and functional studies of microbiome data (Ju and Zhang, 2015). Using the

clustering approach, taxa or samples can be partitioned based on their similarities and dissimilarities. Clustering is a form of unsupervised learning that divides data into meaningful groups, or clusters, based on the information found in the dataset. Similar or related objects are inserted into the same clusters, and dissimilar or unrelated objects are inserted into different clusters. The desire is to get greater intra-class (within the clusters) similarities and inter-class (between the clusters) dissimilarities.

In supervised classification methods, a model is built based on a "training set" which is data from a predefined class. New examples can then be inserted into these predefined groups based on various computational models. The difference between classification and clustering is that in classification objects are inserted into predefined classes based on known information, while in clustering, objects are divided into clusters based on the given data. Three examples of unsupervised clustering algorithms are hierarchical clustering, K-means, and principal coordinate analysis and some examples of supervised classification methods are decision trees, nearest neighbor, naïve Bayes, and support vector machines (Tan et al., 2006).

Both 16S rRNA taxonomy data (composition) and functional gene markers (function) can be used for classification. Functional properties have been proposed to be more discriminatory than compositional properties (Xu et al., 2014), as some communities with the same functions have different compositions. Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) is a tool to predict the function of taxa present in a microbiome (Langille et al., 2013).

There are some software packages that have been developed for performing machine learning on big data. Two examples are Weka and Orange. The Waikato environment for knowledge analysis (Weka) is a suite of machine learning algorithms developed at the University of Waikato, New Zealand. Weka is a workbench with many visualization tools, data analysis algorithms, and predictive modeling. Weka supports many standard data mining tasks including data preprocessing, clustering, classification, regression, visualization, and feature selection (Witten et al., 2011). Orange is another open source machine learning, data mining, and analysis software platform maintained and developed by the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana, Slovenia (Demsar et al., 2013). It has a visual programming front-end for explorative data analysis and visualization. The Orange components are called widgets and are modules for data visualization, subset selection, preprocessing, empirical evaluation of learning algorithms, and predictive modeling. Widgets offer additional functionalities such as data visualizing tools, feature selection, training predictors, comparing learning algorithms (Demšar 2013).

2.6.1 Supervised classifiers

Supervised classification uses examples that are assigned to predefined classes. Two types of datasets are used in supervised classification, training datasets and test datasets. A training dataset is a set of data with known class labels and a testing dataset consists of data with unknown class labels.

Many different classifiers, such as decision trees, nearest neighbor, naïve Bayes, artificial neural networks, and support vector machines (SVM) are in general use. Each

classifier recruits a learning algorithm to make a model that is the best fit for the attribute (e.g. OTU) set. The model performance can be evaluated by preparing a confusion matrix that shows correctly and incorrectly classified data. To compare models, performance metrics, such as accuracy and error rate, are used. Accuracy is the percent of correct predictions of the model, and the error rate is the percent of incorrect predictions of the model. Thus, models with relatively high accuracies and low error rates are desired.

Two types of errors can occur using the above classification methods, training and generalization errors. Training errors are the number of misclassified errors in the training data and generalization errors are the errors seen on previously unseen data during testing. A good model has low rates of both training and generalization errors. If a model works well on training data, but not on test data, it is overfitted. "Overfitting" occurs when a model weights the attributes on just the training data rather than learning to generalize from it. There are various techniques to avoid overfitting, for example using cross-validation.

The performance of different classifiers can be compared to determine which classifier works best for the desired dataset. Using only the accuracy to compare different classifiers may not be reliable because, based on the size of the dataset, the differences observed between the classifiers may not be statistically significant. Therefore, other parameters, such as the receiver operating characteristic (ROC) curve can be used to compare classifiers (Tan et al., 2006).

We describe some of the many classification methods that can be used for microbiome data below.

Decision tree: the decision tree classifier is a hierarchical structure that consists of nodes and edges. "Hunt's algorithm" which is a greedy algorithm[1] is the basis of many decision tree induction algorithms such as ID3, CD4.5, and CART. In this algorithm, training datasets are recursively partitioned into successively purer subsets as shown in Figure 8. A decision tree is a popular method to create and visualize predictive models and algorithms, as it is uncomplicated, easy to understand, and usually effective (Tan et al., 2006).

[1] Greedy learner: use training data to make a classification model before receiving test data.

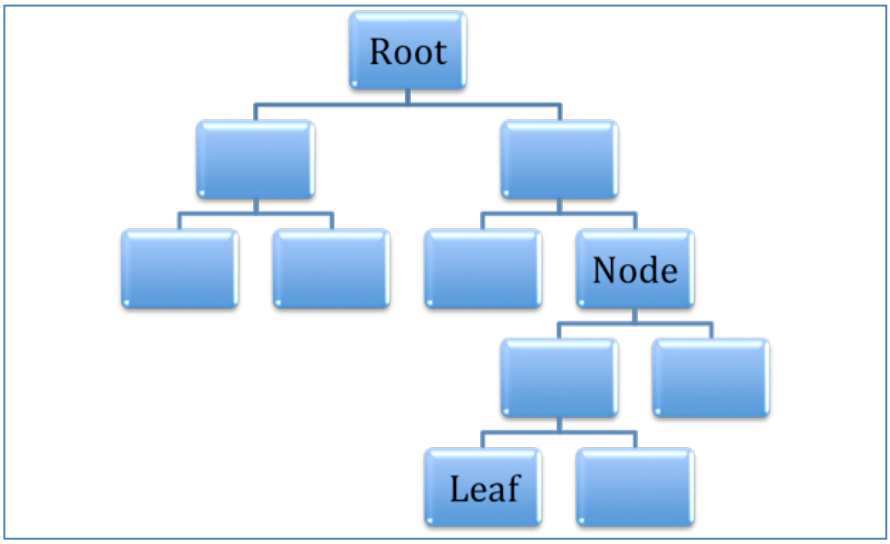


Figure 8 Decision tree and its nodes.
It has a hierarchical structure.

Nearest-neighbor: this approach is a lazy learner[2] as it finds all the training examples that are relatively similar to the test dataset examples (nearest neighbors) and uses this metric to determine the class label. The algorithm computes the distance or similarity between each of the records of the test dataset and all examples of the training dataset to detect the nearest neighbor list. After finding the nearest neighbors, the class label of the test examples is introduced. In this algorithm, the outlying training examples have less effect in the class assignment of the test set.

In the k-nearest neighbor approach, k examples of the training dataset are considered for finding the nearest neighbor. Therefore, choosing a very small k increases

[2] Lazy algorithm: store training data and do not start making classification model until it receives test data.

the chance of overfitting while choosing a large k can cause misclassification due to the presence of very far neighbors.

Lazy learners, such as the nearest neighbor, do not require model building but merely need a proximity measure to find the similar neighbors and assign the class label. However, this method can be expensive because of the need to check the proximity of each test example with all training examples. On the other hand, because greedy algorithms, such as the decision tree algorithm, spend much time on model making, the classifying step itself is relatively rapid. Choosing a suitable proximity measure and data preprocessing are necessary for the nearest neighbor learner. Otherwise, incorrect predictions may be introduced (Tan et al., 2006).

Bayesian classifier: the naïve Bayes classifiers are built according to Bayes' theorem with independent assumptions between predictors and are especially useful when the input dimensionality is high (Tan et al., 2006). Despite the simplicity of this classifier, it often outperforms other more sophisticated classification approaches. Naïve Bayesian is called "naïve" because it assumes that attributes are conditionally independent of each other. This means that the impact of an attribute value on a specified class is not dependent on the values of the other attributes. This assumption reduces computational costs (Tan et al., 2006).

Bayes' theorem is a statistical principle for merging prior knowledge of classes with new evidence that is gathered from input data. This theorem is a useful tool for calculating conditional probabilities. Bayes' theorem is a way of understanding how the probability that a theory is true is affected by new evidence.

The Bayes' theorem formula is:

$$P(T|E) = \frac{P(E|T) \times P(T)}{P(E|T) \times P(T) + P(E|\neg T) \times P(\neg T)}$$

Equation 3 Bayes' theorem formula

Where:

T = the theory (hypothesis) that we want to analyze

E = the new piece of evidence that appears to confirm or reject the theory

P(T) is defined as the *prior probability* of T: our best estimate of the probability of the theory we are analyzing before taking into account the new piece of evidence.

P(T|E) is the probability that T is true given that E is true. It is the *posterior probability* of T. P(T|E) represents the probability that is assigned to T *after* considering the new piece of evidence, E.

To calculate P(T|E), in addition to the prior probability P(T), we require two further conditional probabilities indicating how probable our piece of evidence depends on whether our theory is true (P(E|T)) or not true (P(E|~T)), where ~T is the proposition that T is false.

We want to determine whether the probability that T is true assuming the new piece of evidence is true. This is called conditional probability, the probability that one hypothesis is true provided another be true. For example, when a random card is drawn from a deck of 52, the probability that the card is a jack, P(J) is 4/52 because four jacks are on the deck. However, if we know the card is a face card, the probability the card is a jack

will be 4/12 because 12 face cards are on the deck. This is an example of conditional probability as P(J|F), meaning the probability the card is a jack *given that* it is a face card.

A naïve Bayes classifier estimates the class-conditional probability by assuming the conditional independence of attributes. The conditional independence assumption can be presented as below when the attribute set X consists of d attributes ($X = \{X_1, X_2, \ldots, X_d\}$):

$$P(X|Y = y) = \prod_{i=1}^{d} P(Xi|Y = y)$$

By assuming conditional independence, we just calculate the conditional probability of each X_i, given Y instead of estimating the class conditional probability for every combination of X. This method is practical because it does not need a large training dataset to get an accurate estimate of probability.

To classify a test record, this classifier will calculate the posterior probability for each class Y:

$$P(Y/X) = \frac{P(Y)\prod_{i=1}^{d} P(Xi|Y)}{P(X)}$$

Since P(X) is fixed for every Y, it is sufficient to select the class that maximizes the term, $P(Y)\prod_{i=1}^{d} P(Xi|Y)$ (Tan et al., 2006).

Artificial neural network (ANN): the development of this classifier was influenced by the biological neural system. However, it is simpler than the neural network of a biological system (Sayad, 2011). Many interconnected nodes are found in this type of classification, similar to what we see it in the brain. An ANN includes artificial neurons or so-called nodes and these nodes are connected to each other with different levels of strength or weighting. Stronger connections are given a higher weight value than weaker

connections. A transfer function is built into each node's design. Three types of nodes exist in the ANN model; input, hidden, and output nodes. The input nodes are the starting nodes that bring in the attribute information in the numeric form. The nodes have numbers that reflect their activation levels and a node with more activation has a bigger number than a node with less. This information transfers through the nodes of the network until it reaches the output node and is then presented in a meaningful way to the user (Sayad, 2011). Many ANN models are used for classification. Two examples are the perceptron and multilayer artificial neural networks.

Perceptron: The perceptron is the simplest ANN model, consisting of two types of nodes, input nodes, which code for the input attributes, and output nodes, which represent the output of the model as depicted in Figure 9.

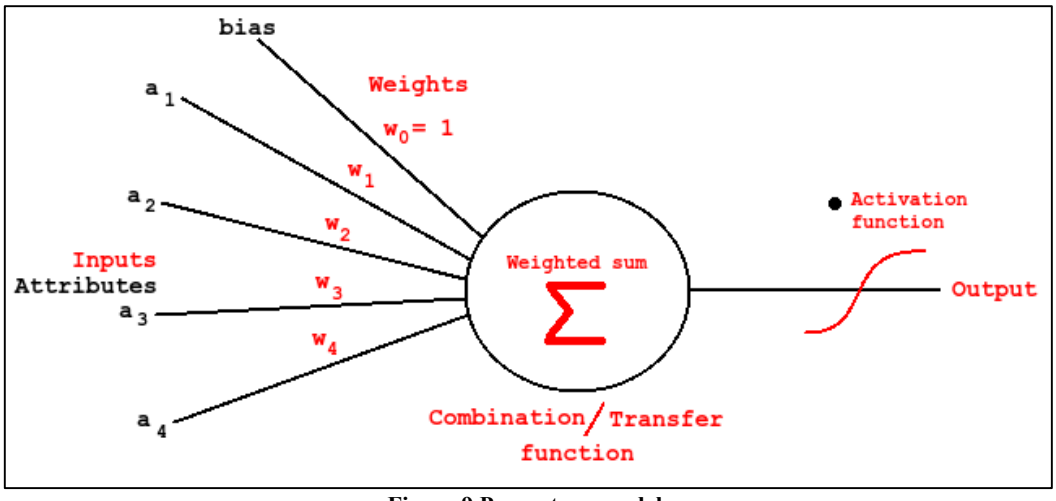


Figure 9 Perceptron model.
This model has input and output nodes that are connected by a weighted link.
https://andynor.net/blog/archive/2013/2/

The input node is connected to the output node by a weighted link which imitates the synaptic connection strength between neurons in the biological system. Training of a perceptron model is similar to what happens in a real biological system which amounts to adapting the weight of the links until they optimize the input-output relationships of underlying data. The weight parameters are adjusted until the output data of the model is consistent with the output of the training examples.

Multilayer artificial neural network: The structure of this model is more complicated than a perceptron in many ways. These include having intermediary layers, also called hidden layers, hidden nodes, and uses more complicated functions as depicted in Figure 10.

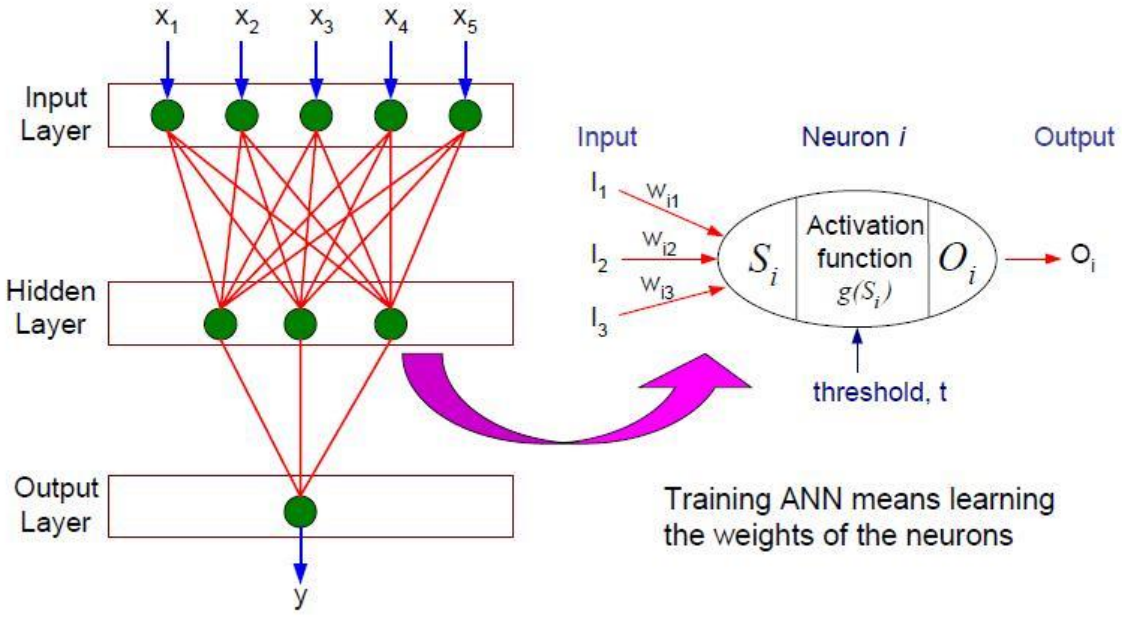


Figure 10 Multilayer artificial neural network.
This model has hidden layers and hidden nodes, and it uses activation functions such as linear, sigmoid, or hyperbolic tangent. http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/ann.html

The time-consuming portion of the ANN method is the training set processing, mainly when the number of hidden nodes is large. However, the classification of testing examples is fast (Tan et al., 2006).

Support vector machine (SVM): SVM originated from statistical learning theory and is effective in many aspects and it handles high-dimensional data well. This approach represents decision boundaries using examples of a training dataset to determine support vectors (hyperplanes).

Many possible hyperplanes separate two high dimensional datasets from each other to form separate classes. The classifier should choose one of these hyperplanes as the separator based on its effect on the testing dataset. Even if the hyperplane separates the training data entirely, we cannot be sure it will work well on the test dataset. Each decision boundary is associated with a pair of parallel hyperplanes that touch the closest examples of each of the classes. The distance between these two parallel hyperplanes is called the margin. The longer the margin, the better the generalization error for that classifier. The classifiers with short margins are more prone to overfitting than those with longer margins and generalize weakly on test datasets as shown in Figure 11.

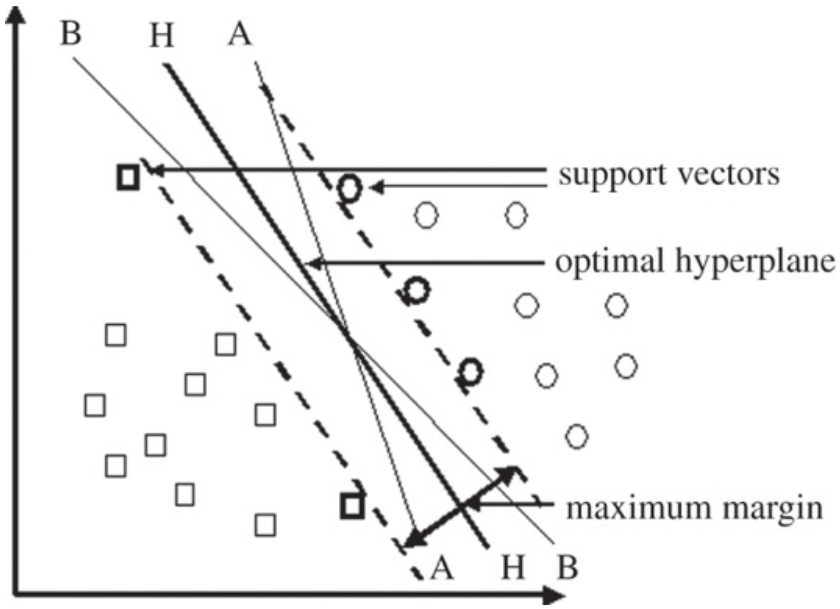


Figure 11 Support vectors scheme.
Support vectors are examples from the dataset that define boundaries between classes. A hyperplane with more extended margin will represent a better classifier.
Ref: http://rsif.royalsocietypublishing.org/content/9/73/1934

Linear SVM: Linear SVM, or maximal margin classifier, is a classifier that seeks a hyperplane that gives the longest possible margin.

Soft margin approach: The soft margin approach is a method that looks for a model that can tolerate small training errors. Therefore, even if two classes are not linearly separable, this method allows the SVM to make linear decision boundaries by accepting some training errors. The learning algorithm should take into account the tradeoff between the margin distance and the number of training errors to make a boundary decision.

Nonlinear SVM: The easiest way to separate two groups is to use linear decision boundaries, including a straight line (one dimension), a flat plane (two dimensions), or an N-dimensional hyperplane. However, for some datasets, a nonlinear SVM may differentiate two groups more efficiently than a linear SVM.

In this method, the SVM will transform the original data from its original coordinate space into a new space using a nonlinear kernel function in such a way that a decision boundary can be defined to separate classes in the new space (Tan et al., 2006) as illustrated in Figure 12. This is called the "kernel trick," which means that a nonlinear function transforms the data to a higher-dimensional level to make linear separation possible (Sayad, 2011).

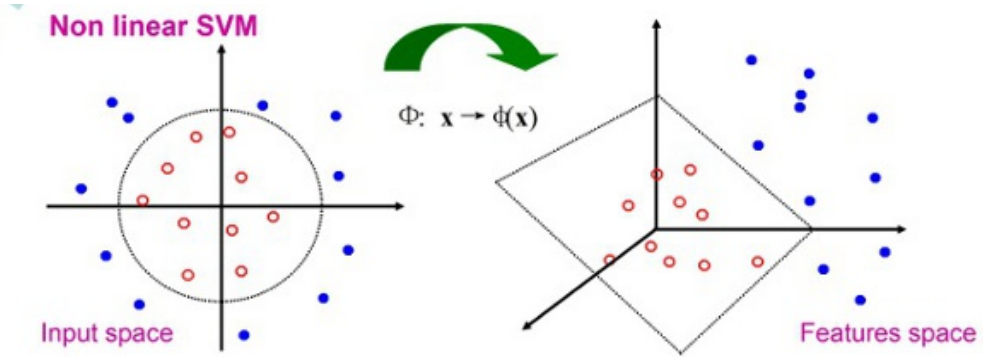


Figure 12 Nonlinear support vector machines.
The data will transform to a higher dimensional level to make linear SVM possible.
Ref: http://www.intechopen.com/books/air-pollution/artificial-neural-networks-for-pollution-forecast

Ensemble methods: to increase the accuracy of classification, the predictions of several classifiers may be combined as shown in Figure 13. This approach is called the ensemble classifier or classifier combination. It makes a series of individual base classifiers from the training dataset, and then by taking a vote on the resulted predictions, performs the classification.

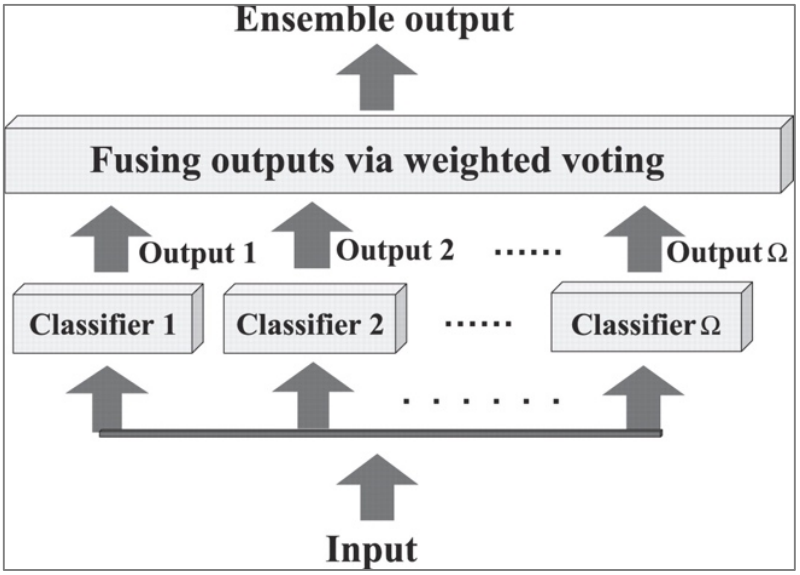


Figure 13 Ensemble classifier.
It is a combination of several classifiers.
Ref: http://bioinformatics.oxfordjournals.org/content/22/14/1717/F2.expansion.html

Many ways exist to make ensemble classifiers, including manipulating the training dataset, input features, class labels, and learning algorithm (Tan et al., 2006).

Random forest: The Random forest classifier is an ensemble method specific for decision tree classifiers. It aggregates the predictions produced by different decision trees. The differences between the decision trees depend on the values of their random vectors. As shown in Figure 14, the random vectors could be generated from a fixed probability distribution (Tan et al., 2006).

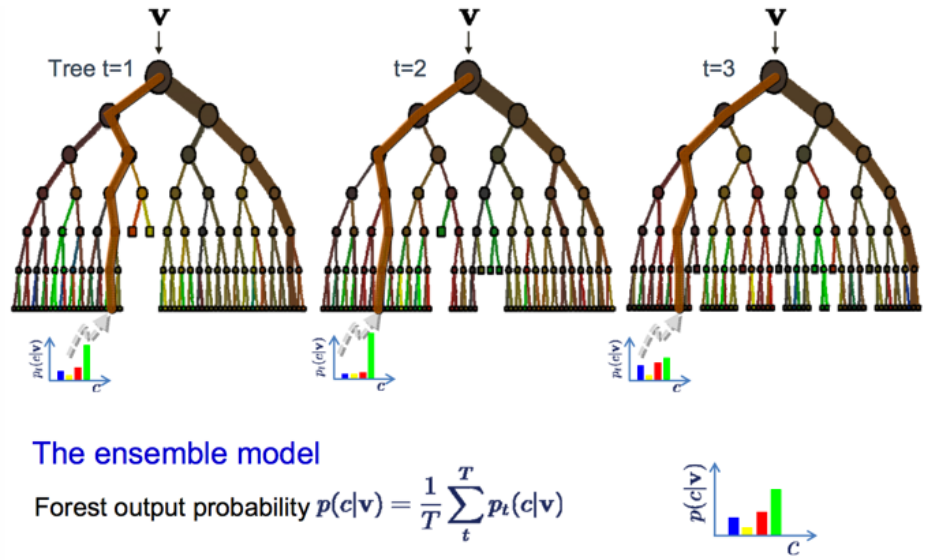


Figure 14 Random forest classifier.
It aggregates the predictions produced by different decision trees.
Ref: http://stackoverflow.com/questions/17031056/using-c4-5-classifier-with-multiple-outcomes

2.6.2 Classification validation and classifier performance evaluation

There are many methods to compare the performance of a classifier including the holdout method, random subsampling, cross-validation, and bootstrap. In this study, the results of cross-validation were reported. In the cross-validation approach, the input data is partitioned into several segments with equal size. For each run, one of the portions is employed as the test dataset while the others are used as the training dataset and this process continues until all of the partitions are used precisely once for testing. The training dataset is used to generate the classification model and the test dataset is used to evaluate the classifier performance. The portion of the data that is selected for each of the training and test datasets is at the discretion of the user. For example, it can be divided into a 3-fold cross validation set or $^{2}/_{3}$ training to $^{1}/_{3}$ test data. The sum of the errors from each run is the total error of this method and the accuracy is calculated based on the result of applying

this classifier on all the test datasets. For the test dataset, each record is recruited only once, and the remaining data is used for training as demonstrated in Figure 15. There are many advantages for this method including a) using the maximum possible data for the training dataset; b) the test dataset uses the whole dataset, and c) the training and test datasets are mutually exclusive. The disadvantages can be the high variance of the performance and the high computational expenses.

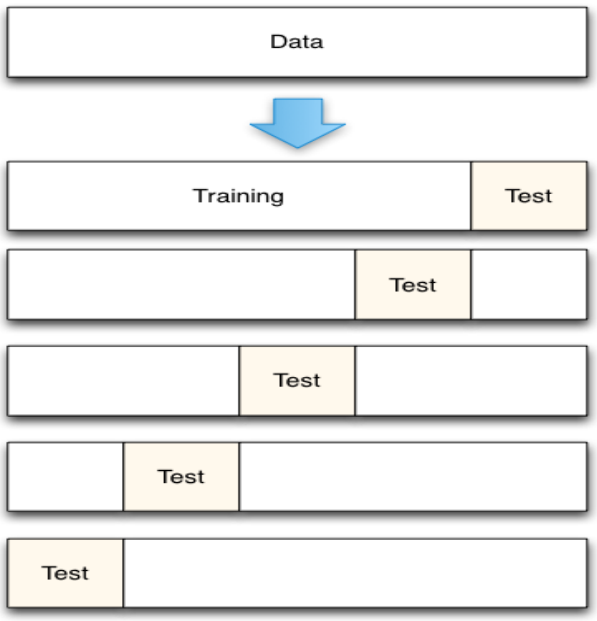


Figure 15 Cross-validation.
The input data is partitioned into k segments with equal size. For each run, one of the portions is employed as the test dataset, other are used as the training, and this process continues until all of the partitions is used exactly once for testing. Ref: http://scott.fortmann-roe.com/docs/MeasuringError.html

Another performance measure for classification problems is error rate or accuracy. However, higher accuracy does not necessarily imply better performance on target data and it is recommended to use multiple measurements to check the performance of a

classifier, such as an accuracy, sensitivity, specificity, and area under ROC curve as no single measure is a perfect evaluator of a classifier. Sensitivity is a measure of how well a binary classifier correctly detects a condition or probability of correctly labeling a target class member. Low sensitivity indicates a high false-negative rate and is a weak classifier to rule out the disease class. On the other hand, specificity is the statistical measure of how well a binary classifier correctly detects the negative cases.

The receiver operating characteristic curve (ROC) is a graphical plot to illustrate the performance of a classifier and represents the trade-off between true positive rate (TPR) and false positive rate (FPR). The ROC curve is made by plotting TPR (y-axis) against FPR (x-axis) as shown in Figure 16. The TPR is also called sensitivity, and 1-FPR is also known as the specificity. Each point on the curve belongs to one of the models that are produced by the classifier.

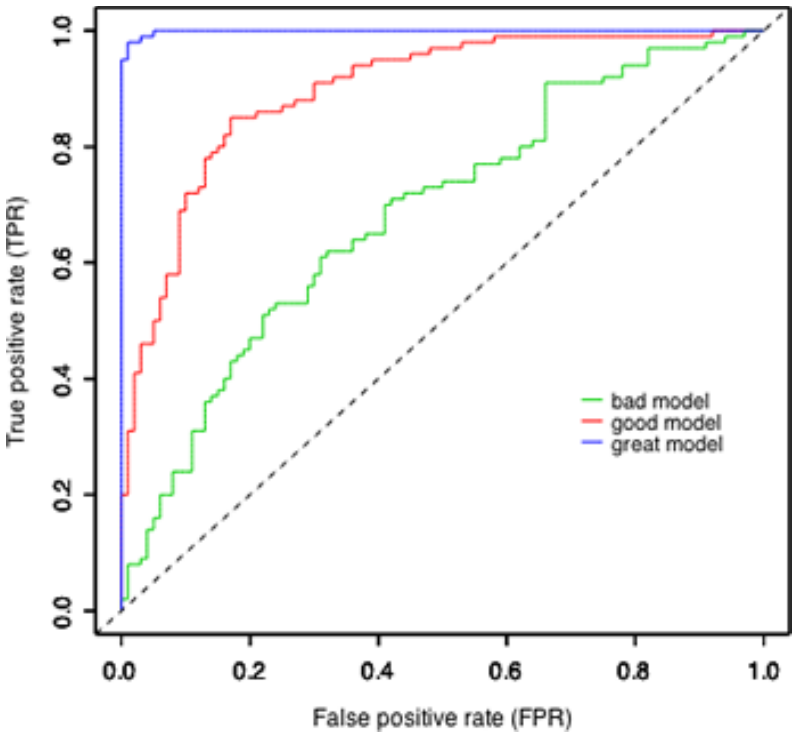


Figure 16 The receiver operating characteristic curve (ROC).
The ROC curve is a graphical plot to illustrate the performance of a classifier and represent the trade-off between true positive rate (TPR) and false positive rate (FPR).
Ref: http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture22.htm

A good classifier is located on the upper left as much as possible. A model that is close to the main diagonal is a model that makes a random guess and not a useful model.

The ROC curve is very helpful for comparing classifiers to each other and differentiating their relative performance. In addition, the area under the ROC curve (AUC) can also be used for comparing classifiers. A perfect model will show AUC=1 and a model with random guess will represent AUC=0.5. Therefore a better model will have a higher AUC (Tan et al., 2006). A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

- 0.90-1 = excellent
- 0.80-0.90 = good
- 0.70-0.80 = fair
- 0.60-0.70 = poor
- 0.50-0.60 = fail

2.7 Application of classification in microbiome studies

In the last few years, there have been attempts to use classification to extract useful information from microbiome data. Some studies have used SVM, ANN, or ensemble classification methods for microbiome analyses and we discuss them below.

Nakano et al. reported that SVM, ANN, and a decision tree helped them to classify the oral microbiota and malodor microbiome in saliva and these classification methods proved to be useful for screening saliva for oral malodor before visits to specialist clinics (Nakano et al., 2014).

Yemin's group combined SVM classification with functional feature selection to identify age-related functional characteristics in metagenomes collected from the human gut. They showed that the combination of feature selection with SVM yields biologically meaningful results and simplified age classification of new human gut metagenomes (Yemin et al., 2013).

Wisittipanit et al. used SVM and KNN classification to distinguish samples from patients with Crohn's disease and ulcerative colitis from healthy control samples. Using these methods, the authors reported OTUs or microbial species that were differentially

abundant between patients and healthy controls at specific intestinal locations (Wisittipanit et al., 2015).

Schubert et al. reported microbiome alterations that potentiated *Clostridium difficile* infections after antibiotic use. They built a random forest regression model to predict *C. difficile* colonization levels based on microbial relative abundance data. Using this model, they identified *C. difficile*-related bacteria that were colonization resistant. Interestingly, they were unable to find these distinctive bacterial groups using other correlation approaches (Schubert et al., 2015).

In a mouse model, a random forest regression modeling approach predicted the number of tumors present at the end of the study based on the original bacterial composition of the mouse microbiome (Zackular et al., 2015).

3 MATERIALS & METHODS

3.1 Metadata and datasets

Metadata and 16S rRNA sequence datasets from previously published studies were used in this project, and they were analyzed by three OTU selection approaches. Two studies were selected as benchmarks for preliminary hypothesis analysis: Zackular et al. 2014 (referred here as Benchmark-1) and Wu et al. 2013 (referred here as Benchmark-2). A summary of these selected studies on CRC and adenoma microbiome are listed in Table 3. They were chosen as they were publicly available and similar in design to our own study (MBO1) which compared specimens collected from subjects with colorectal cancer, benign adenomas, and normal colonoscopy results. In addition to the raw sequences, we needed the clinical metadata to define the disease states. For other similar studies, this information was not available, and request for this information from the corresponding authors was unsuccessful.

There were also a few studies that were available with the necessary information for assigning samples, but the quality of the reads was very low as most of the reads were discarded at the preprocessing steps. As such, we decided not to include them as the low quality of the sequences could severely affect the OTU clustering and classification methods. Additionally, there were also some studies with insufficient numbers of subjects that we did not use because of the reduced reliability of results due to the small sample size.

The DNA sequencing technology used and the variable region analyzed for these two benchmarks were different. Benchmark-1 used the Illumina MiSeq platform and V4 variable region of 16S rRNA, whereas Benchmark-2 utilized the 454 Roche (pyrosequencing) technology and V3 variable region. On the other hand, the sequencing method for the MBO1 polyp data was the Ion-Torrent PGM and the primers of V1-V2 were selected for analysis. As the technology and primers used in these studies are different from each other, we tried to choose parameters for the sequence analysis pipeline that can work efficiently for all of these methods. For example, with the 454 (pyrosequencing) and Ion-Torrent systems there is a high chance of homopolymer errors, whereas, with the Illumina system, the rate of ambiguous bases is greater. Therefore, in the preprocessing steps, we checked sequences for both of these errors to make sure that the pipeline worked well for all of these datasets.

Table 3 Information of the two studies selected as benchmarks.
Information about their subjects, DNA extraction, and sequencing technology are summarized here.

	Benchmark-1: Zackular et al., 2014	Benchmark-2: Wu et al., 2013
Type of samples	Fecal sample	Fecal sample
Disease	CRC & Adenoma	CRC
Population	Canada-USA	China
Sample size	30 CRC, 30 adenoma, 30 healthy controls	20 CRC, 20 healthy controls
DNA extraction	Power- Soil-htp 96 Well Soil DNA Isolation Kit	QIAamp DNA Stool Mini Kit
Sequencing machine	Illumina MiSeq	Genome Sequencer FLX System (Roche) pyrosequencing
rRNA variable region amplified	V4	V3
Number of reads	12,180,024	727,860
Data depository	http://www.mothur.org/MicrobiomeBiomarkerCRC	https://www.ncbi.nlm.nih.gov/sra/SRX152609[accn]

The MBO1 polyp and healthy control dataset was collected for a clinical trial study sponsored and funded by Metabiomics Corp. (Sterling, VA, USA) and conducted at the Metropolitan Gastroenterology Group (Chevy Chase, MD, USA) and George Mason University, Microbiome Analysis Center (Manassas, VA, USA) under an IRB approved clinical trial protocol described on clinicaltrials.gov (ID# NCTO2141945). The biopsy (BS), rectal swabs (SS), and home stool swabs (HS) were collected from subjects that had undergone routine colonoscopy for polyp detection. In total, 552 samples were collected, including both polyp positive (n=316) and polyp negative (n=236) subjects. Some of the

subjects did not submit all types of specimens, which resulted in the collection of 231 rectal swabs (SS), 183 home stool swabs (HS) and 138 Biopsies (BS) that were used as material inputs for sequencing (Table 4). Samples were kept at -20°C in RNALater until DNA samples were extracted using FastDNA Spin Kit for Soil (MP Biomedicals, Solon, CA, USA). The sequencing was performed using V1-V2 bacterial primers and Ion Torrent Personal Genome Machine (Life Technologies, USA) located at the Microbiome Analysis Center, George Mason University. The number of reads collected from the machine for the 552 samples was 12,646,278.

A summary of polyp metadata information is shown in Table 4. The average age of subjects was 62 years old and average BMI was 27. Participants were 131 (60%) male and 87 (40%) females. The ethnicity of subjects was African-American (21, 10%), Caucasian (192, 88%), and Asian-American (4, 2%). Based on colonoscopy results, subjects were categorized into two groups: polyp-negative (polyp-N) and polyp-positive (polyp-Y).

Table 4 MBO1 polyp dataset information.
All of the 552 samples were collected in Washington DC metro area. Three type of samples were collected. However, not all of the subjects have all three samples. After the colonoscopy, subjects were divided into two groups of polyp-Y (ones with detected polyp) and polyp-N (ones without polyps). DNA extracted from all samples with the same DNA extraction kit and sequenced in Microbiome Center of George Mason University.

Polyp Study	
Type of samples	Rectal swab (SS), home stool swab (HS), biopsy (BS)
Disease status	Patients with colon polyps and healthy control individuals
Population	USA
Groups	Polyp positive (polyp-Y) and polyp negative (polyp-N)
Sample size	polyp-Y (316); polyp-N (236) [231 swabs, 183 stool samples, 138 biopsies]
DNA extraction	FastDNA Spin Kit for Soil (MP Biomedicals)
Sequencing approach	Ion Torrent Personal Genome Machine
rRNA variable region	V1-V2
Number of reads	12,646,278

3.2 Sequence analysis

There are many different parameters that can be selected for the 16S rRNA sequencing analysis pipelines. For example, preprocessing steps can be very different. Decisions made for selecting preprocessing parameters, clustering criteria, and sample depth cut-offs all affect downstream analyses. There should be a balance between removing low-quality sequences and keeping enough sequences for a statistically significant depth of coverage. Highly conservative preprocessing may remove real sequences and highly permissive processing may let noisy data distort or overwhelm real community patterns. The steps of the optimized pipeline are preprocessing, OTU clustering, and construction of an OTU abundance table. This is then followed by alpha and beta diversity analysis, the performance of statistical tests to find significant OTUs, and finally classification and classification validation. For the polyp dataset, an additional prediction step was performed in which we used the best performing classifier to predict the class of naïve unknown

samples. We developed a unique pipeline to accommodate the different sources of data used in the study which is summarized in Figure 17. The details of each step are presented below.

Preprocessing: preprocessing is the first step of sequence analysis and the results can be very different based on the parameters that are applied. In this study, sequences were processed based on the following criteria:

1- The maximum number of ambiguous bases was 5.

2- The maximum number of homopolymers was 8.

3- All chimeric reads were found and removed using UCHIME. The reference database for chimera removal was RDP.gold (https://rdp.cme.msu.edu) (Release 11, Update 4; May 26, 2015).

4- All undesirable lineages (Mitochondria-Chloroplast-Eukaryotes-Archaea) were removed by using the Naïve Bayes classifier (Wang et al., 2007).

5- Reads with the low-quality score were removed.

6- Singletons (reads that appear just once) were discarded.

7- Reads were trimmed at a fixed length for the UPARSE and UPGMA method. For UCLUST method, reads were sorted based on length.

For Illumina sequences, there was a preliminary step of merging paired-end reads. As there were forward and reverse pair reads, the read pairs were merged to make one read. The merging was performed according to the method described in Edgar & Flyvbjerg 2015.

OTU selection: OTU selection was performed using three methods, UPARSE, UPGMA, and UCLUST. After finding OTUs or centroids, all the reads were mapped to

OTUs using an identity threshold of 97%. Thus, any read with the identity of 97% or higher to a given OTU would be inserted into that OTU cluster. Otherwise, they were assigned to another cluster. An abundance table was constructed that included all the OTUs and their abundance for each sample.

The OTUs were aligned to 16S reference sequence database GreenGenes (http://greengenes.lbl.gov) using the RDP classifier (Wang et al., 2007), and a taxonomic ID was generated for each OTU.

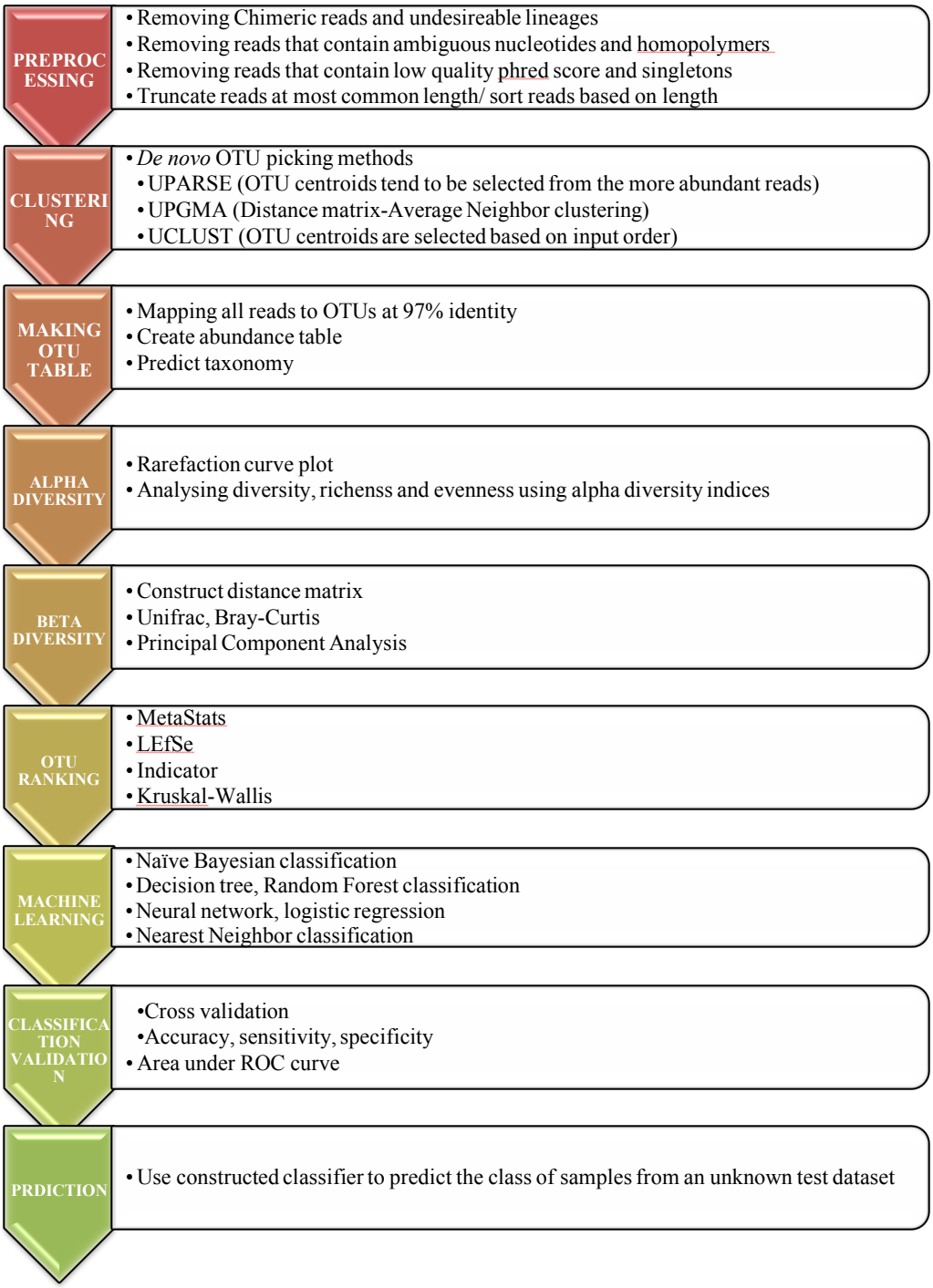


Figure 17 Sequence analysis pipeline optimized for this study.
After preprocessing the sequences and removing unwanted reads, three OTU clustering methods were performed, and three abundance tables were made. Downstream analysis including alpha and beta diversity analysis, finding significant OTUs, machine learning, and prediction is conducted afterward.

3.3 Alpha and beta Diversity analyses

After generating abundance tables, several statistical methods were performed such as alpha diversity measurements of richness and evenness and the identification of the presence or absence of specific taxa (i.e., indicator). Rarefaction curves were drawn to analyze the sequencing depth and then alpha diversity indices such as Shannon, Simpson, invSimpson, and observed species (sobs) were calculated for all of the samples.

Beta diversity was analyzed by generating a phylogenetic tree and this was followed by UniFrac and PCoA analysis. Specifically, the processed reads were aligned using UPGMA algorithm and a phylogenetic tree was produced using FastTree (http://microbesonline.org/fasttree/) (Price et al., 2009; Price et al., 2010) which makes an approximate-maximum-likelihood phylogenetic tree. Then weighted and unweighted UniFrac metrics were calculated using the tree and the results were visualized using principal coordinate analysis.

3.4 Statistical analysis tests

We used Kruskal-Wallis, Metastats, LEfSe, and Indicator to find significantly different OTUs between groups.

1. Kruskal-Wallis (KW) finds OTUs with a significantly different mean rank between groups.
2. LEfSe discovers OTUs with significantly different abundance and biological relevance among groups.
3. MetaStats determines OTUs with significantly different mean proportion and variance among groups.
4. Indicator detects the indicator species of each group.

3.5 Classification methods

Several methods were used to build classification models for all three benchmarks:

1. The naïve Bayesian algorithm that classifies based on posterior probability.
2. The decision tree algorithm that performs a recursive partitioning.
3. The random forest algorithm that aggregates the predictions produced by different decision trees.
4. K-nearest neighbor algorithm that detects the class of test example based on nearest neighbors in the training set.
5. The neural network algorithm that is a multilayer perceptron.
6. Support vector machine algorithm that finds the best hyperplane to separate two groups.

The process of data mining was performed using Orange data mining tool, V. 2.7 (http://orange.biolab.si/), and the pipeline is shown in Figure 18. The classification was performed under two different conditions: once with all detected OTUs as classification features (or attributes) and once with only significant OTUs that were detected by the above statistical methods. The trained classifier obtained using the significant OTUs was then used on test datasets to analyze the performance of the classifier on naïve samples.

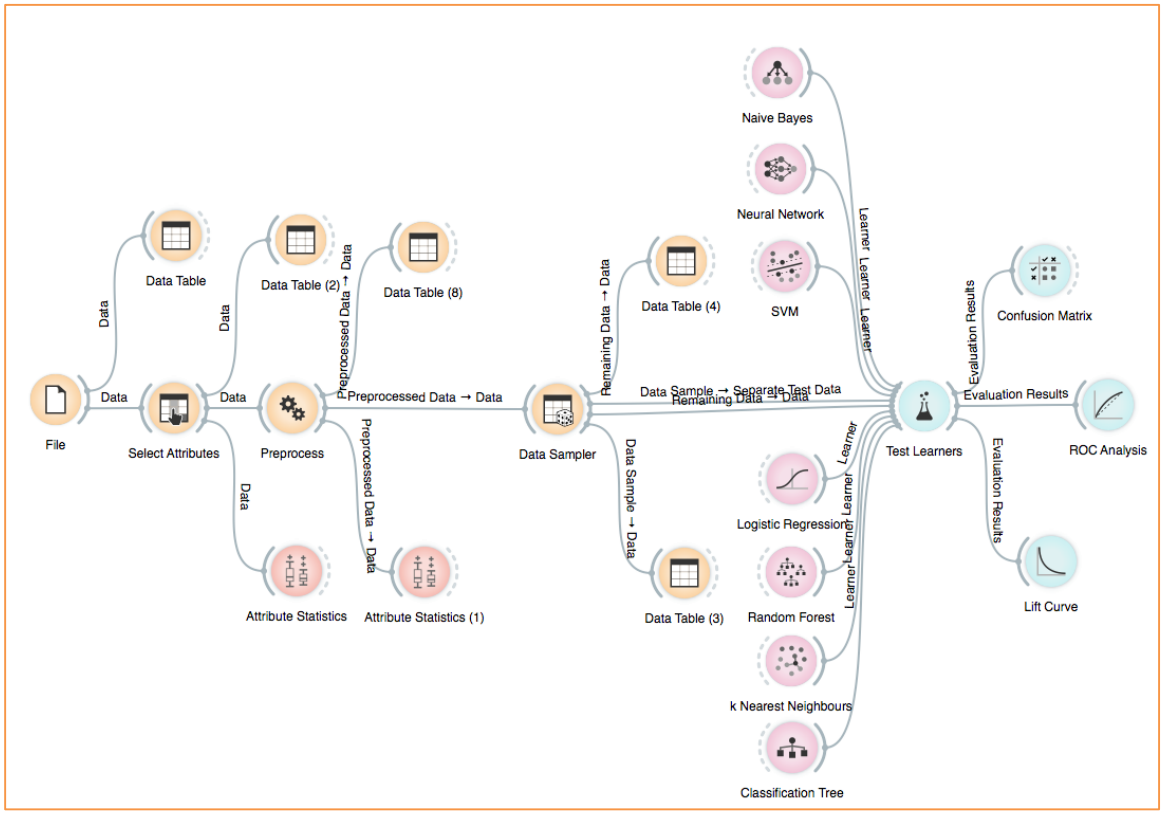


Figure 18 Classification pipeline in Orange data mining tool.
OTU abundance table was uploaded as the input file. The data was partitioned into training and test dataset using 5-fold cross-validation. Training data was used for making the classifier. The resulted classifier then applied to test data to check the performance.

3.6 Classifier validation

The cross-validation method was used as the method of choice for evaluating the each of the classifiers and their classification accuracy, sensitivity, specificity, and area under the curve. Sensitivity, specificity, and classification accuracy were calculated as follows:

Sensitivity= $\frac{TP}{TP+FN}=\frac{TP}{P}$

Specificity= $\frac{TN}{TN+FP}=\frac{TN}{N}$

Classification accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$

Where T= True; F=False; P=Positive; N=Negative.

Orange data mining tool was used for validation as depicted in Figure 18.

3.7 Predictions

The next step is checking the classifier on a separate naïve test dataset to assess the predictive power of these classifiers. A WEKA module was used to generate these naïve test sets along with corresponding training sets. Then the predictions were performed utilizing the orange prediction pipeline shown in Figure 19.

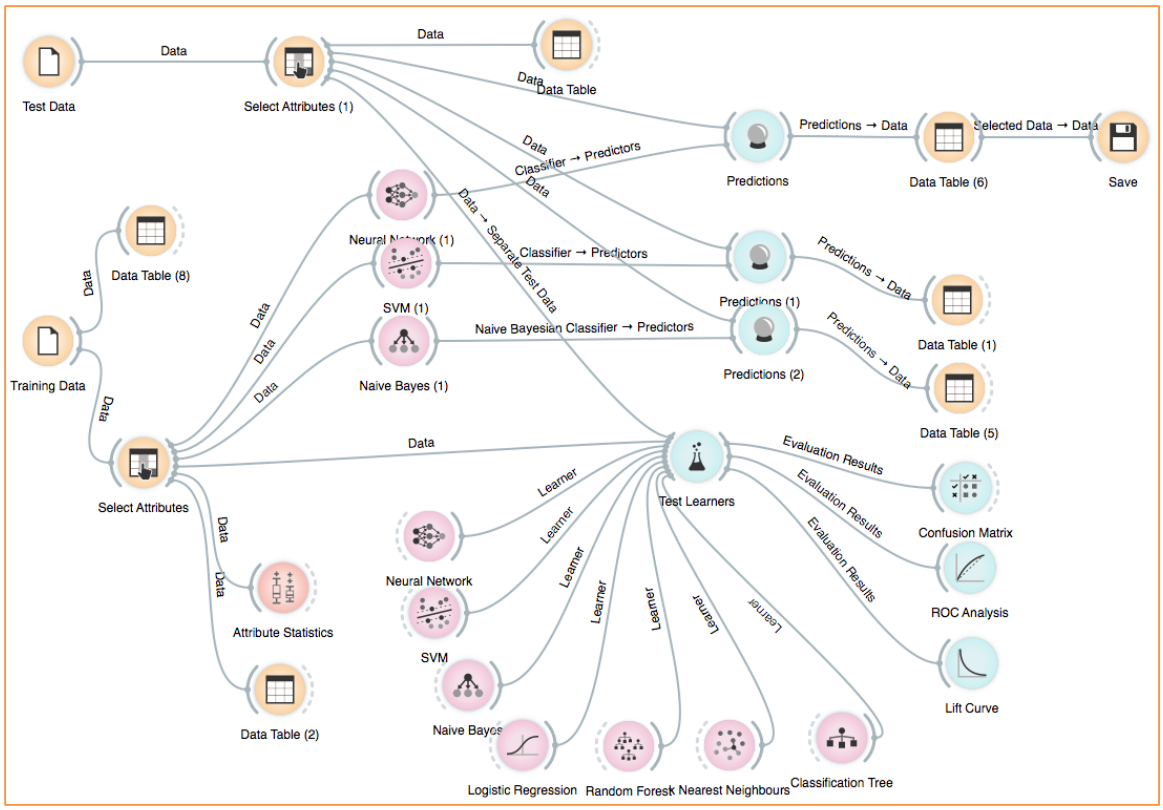


Figure 19 Prediction pipeline in Orange data mining tool.
Classifiers were constructed based on training data, and then they were used to predict the class of a separate test dataset.

4 RESULTS

4.1 Benchmark-1 results

The sequences from Benchmark-1 were preprocessed according to the steps and parameters described in the methods section (Figure 17), and OTUs were clustered using three methods: UPARSE, UPGMA, and UCLUST. The number of OTUs produced by each method were 2560, 4340, and 303184 for UPARSE, UPGMA, and UCLUST, respectively. Thus, UPARSE detected the lowest number of OTUs and UCLUST the highest.

4.1.1 Benchmark-1 rarefaction

The rarefaction plots for the UPARSE analysis of Benchmark-1 are shown in Figure 20. Based on the rarefaction curves, sequencing depth is adequate for all three groups of healthy control, adenoma, and cancer samples. The sequencing depth is adequate as the plots reached a plateau which means that the number of detected species or OTUs would not increase by increasing the number of sequences per sample. As UPARSE generally had better results in our study, only the results of UPARSE are shown here.

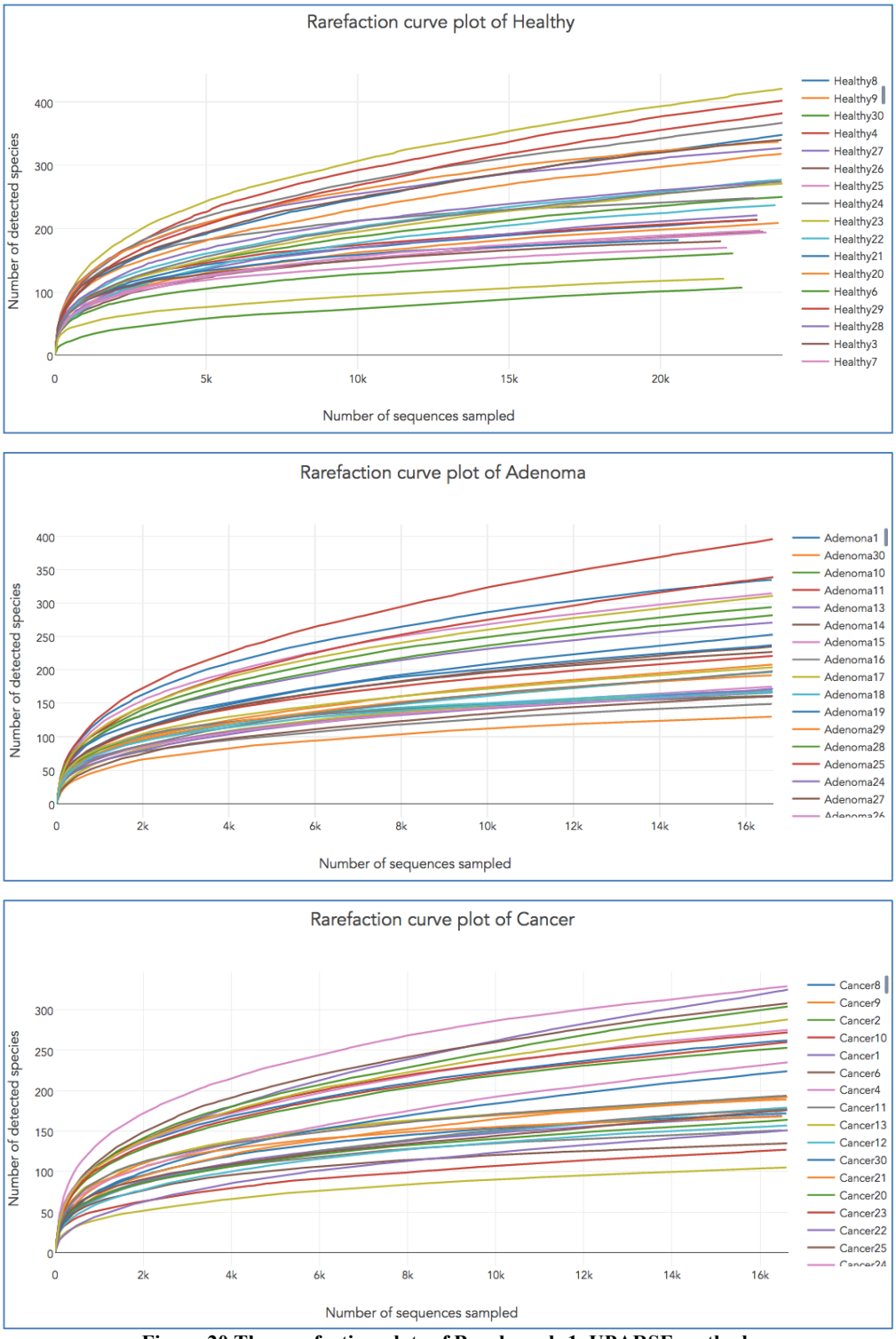


Figure 20 The rarefaction plots of Benchmark-1, UPARSE method.
The sequencing depth is good because all plots reached a plateau which indicates that by increasing the number of sequences the number of detected species would not change.

4.1.2 Benchmark-1 alpha diversity

Alpha diversity results of Benchmark-1 are shown in Table 5. The results of different diversity indices for the three methods of clustering is summarized in this table. Higher Shannon index indicates a more diverse community with a higher richness and evenness. A lower Simpson index means a higher diversity (evenness). The higher the Simpson index is, the less diverse the sample will be. Higher sobs index indicates a higher richness.

Table 5 The Benchmark-1 alpha diversity results.
For each diversity index, the average in that group has been shown. In all three methods, the average diversity of the healthy control group is higher than either the adenoma and cancer groups. nseq: number of sequences.

	UPARSE			UPGMA			UCLUST		
Diversity index	**Average in each group**			**Average in each group**			**Average in each group**		
Groups	Adenoma	Cancer	Healthy	Adenoma	Cancer	Healthy	Adenoma	Cancer	Healthy
nseq	55950	65286	67443	56466	65977	74511	103159	128326	132240
Shannon	3.35	3.34	3.53	3.4	3.4	3.6	4.4	4.4	4.7
Simpson	0.08	0.08	0.06	0.08	0.08	0.06	0.05	0.05	0.04
invSimpson	15.85	16.68	18.66	16.2	17.02	19	24	25.5	28.4
sobs	292	289	321	391.5	396	464.4	5405.2	6707	7387

For all the indices (Shannon, Simpson, invSimpson, and sobs), the average diversity of the healthy control group is higher than either the adenoma and cancer groups. Specifically, the higher Shannon, invSimpson, sobs and lower Simpson in healthy control compared to cancer and adenoma indicates the richness and evenness decreased in the adenoma and cancer groups compared to the healthy control subjects. The reason for lower

diversity in the disease state may be a shift of the bacterial population from a normal diverse community to a few allochthonous bacteria in response to the disease state.

4.1.3 Benchmark-1 beta diversity

To compute differences between microbial communities, the phylogenetic UniFrac metric was used and visualized with PCoA plots. The PCoA plots show that UniFrac metric did not discriminate between groups. Specifically, there was no clear separation between the samples in each group using the first three principal components as shown in Figure 21, Figure 22, and Figure 23. This means that this metric could not differentiate healthy and disease state from each other.

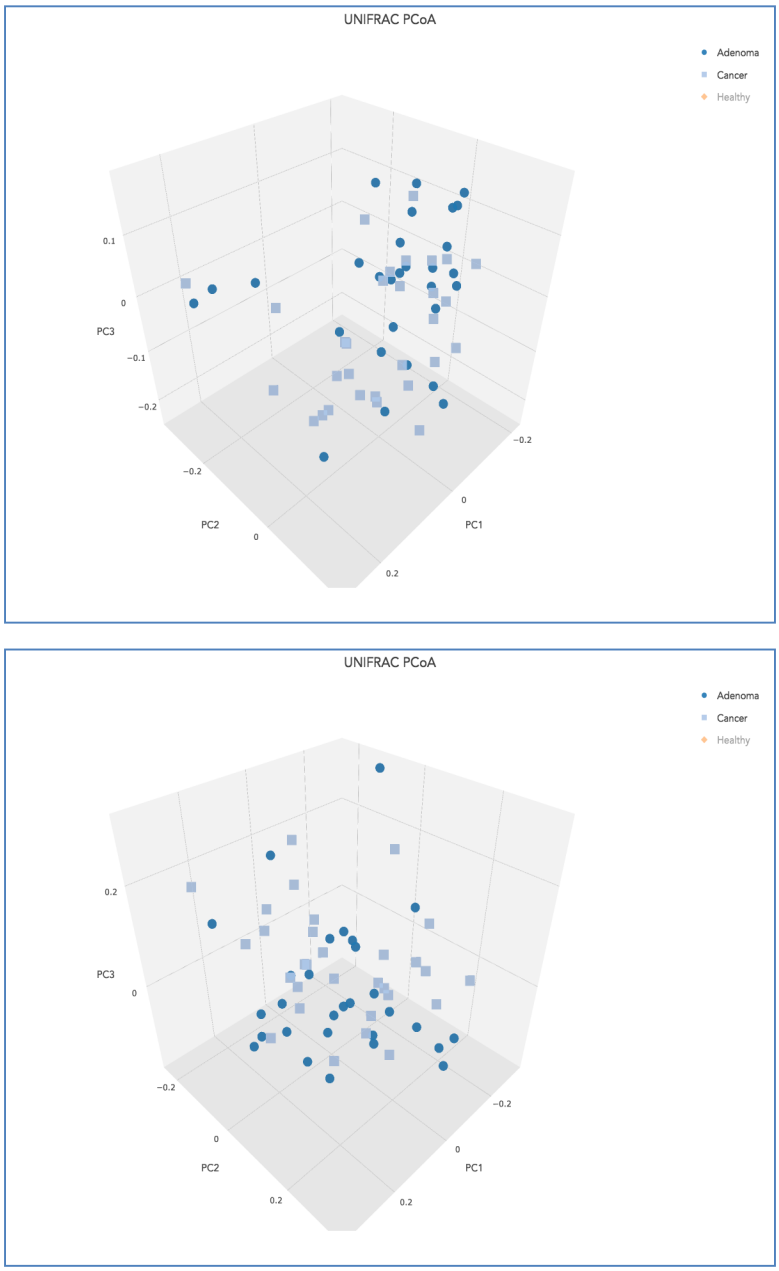
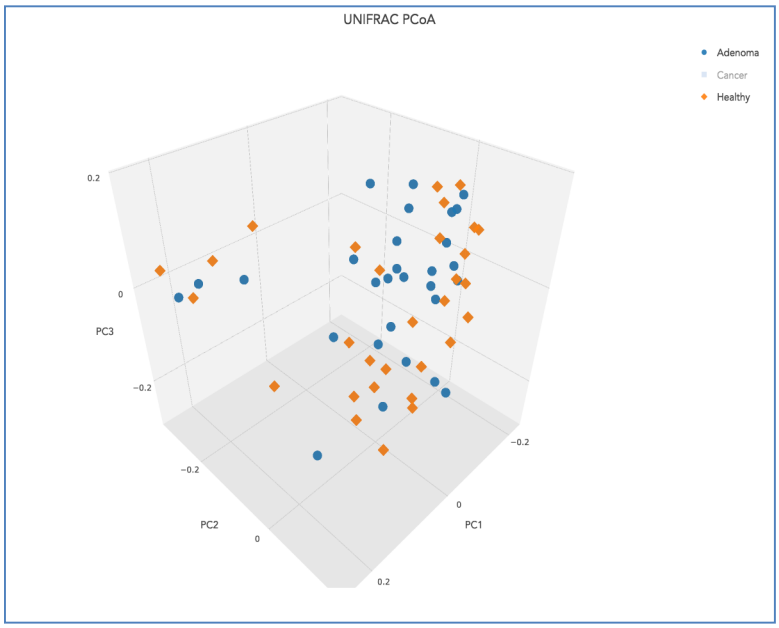


Figure 21 The Benchmark-1 UniFrac PCoA visualization for adenoma and cancer. Neither unweighted UniFrac (top) nor weighted UniFrac (bottom) was successful in differentiating binary groups from each other. The axes are principal components.



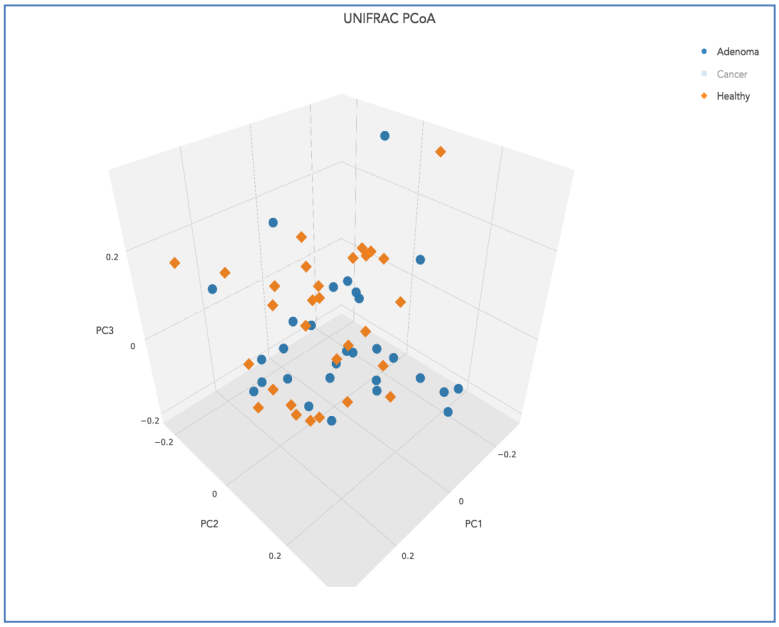


Figure 22 The Benchmark-1 UniFrac PCoA visualization for adenoma and healthy. Neither unweighted UniFrac (top) nor weighted UniFrac (bottom) were successful in differentiating binary groups from each other. The axes are principal components.

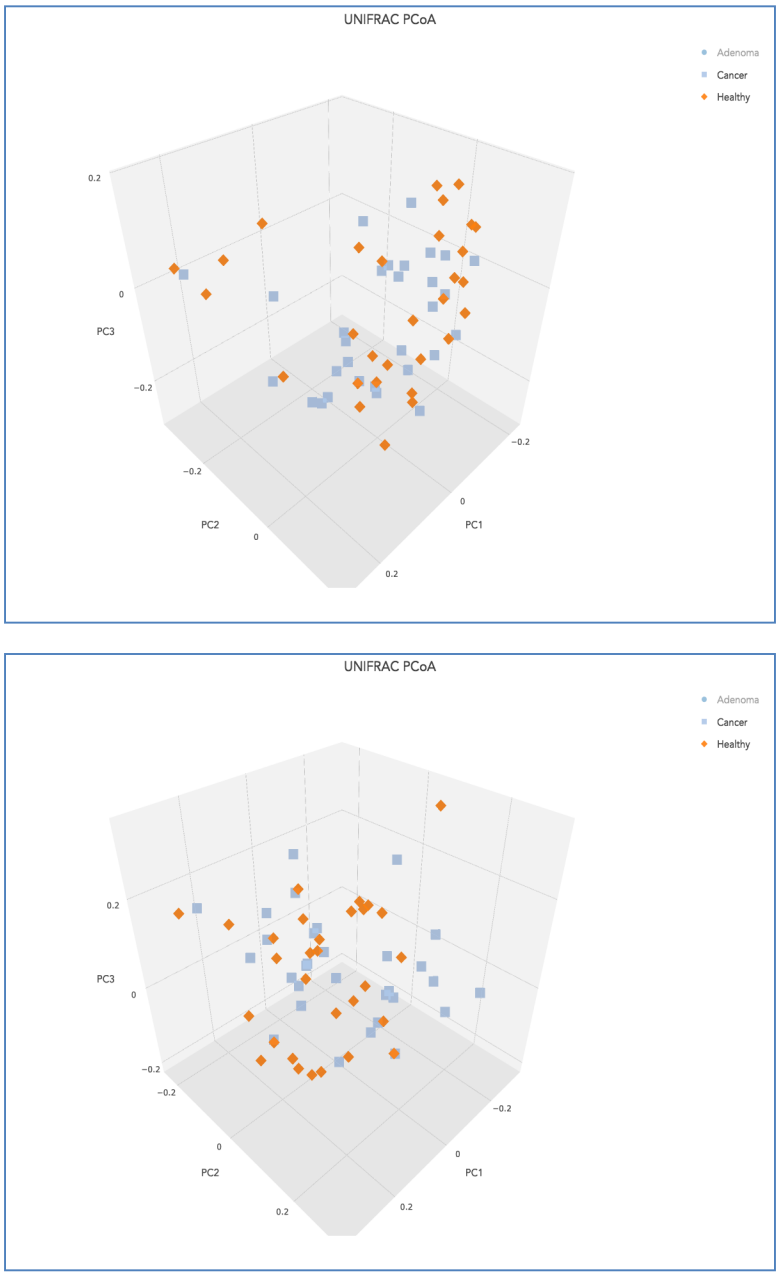


Figure 23 The Benchmark-1 UniFrac PCoA visualization for cancer and healthy. Neither unweighted UniFrac (top) nor weighted UniFrac (bottom) were successful in differentiating binary groups from each other. The axes are principal components.

A Bray-Curtis metric was used as a quantitative non-phylogenetic β-diversity metric. The calculations were performed for the relative abundance of each pair of samples. The PCoA based on the Bray–Curtis dissimilarity matrix revealed that the clustering of intestinal microbiota between each group was similar as showed in Figure 24.

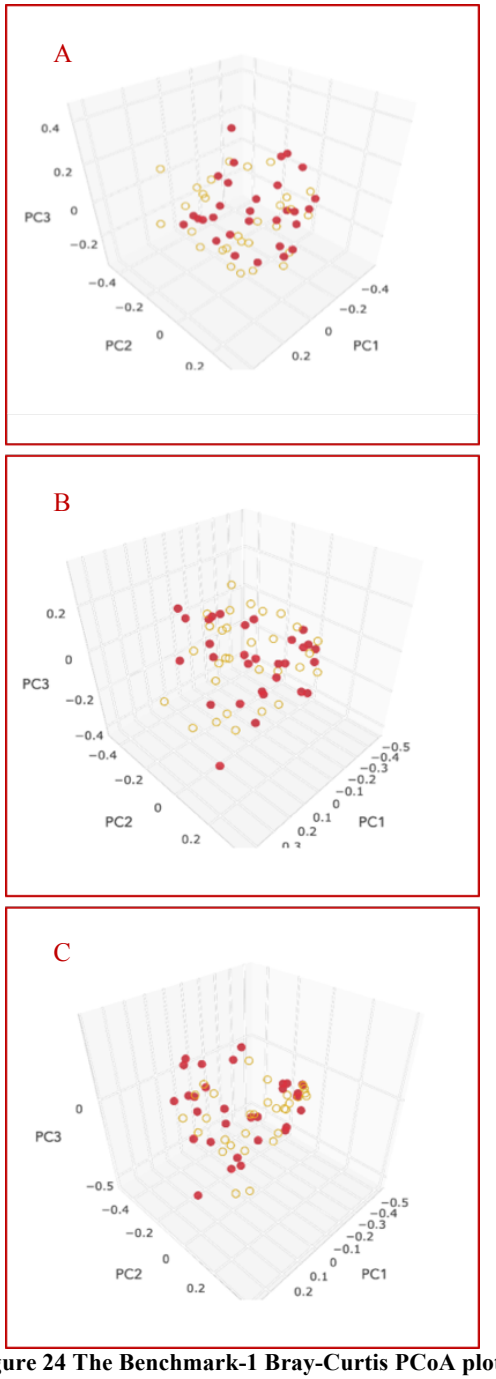


Figure 24 The Benchmark-1 Bray-Curtis PCoA plots.
A: Adenoma(red)-Cancer(yellow), B: Healthy (red)-Cancer (yellow), C: Healthy (red)-Adenoma (yellow). PCoA based on the Bray–Curtis dissimilarity matrix of species abundance revealed that the structure of intestinal microbiota between each of these two groups was similar. The axes are principal components.

4.1.4 Benchmark-1 significant OTUs

We utilized a Kruskal-Wallis, MetaStats, LEfSe, and Indicator to find bacterial species that are significantly different between binary comparisons of the groups of this study. The significance level was 0.05 for all of the tests. Each method enumerated the bacterial species, genera, orders, and families that were differentially present in one group compared to the other. Some of the differentiating taxa were found with more than one of these methods, while some were found by only one of the tests. The Kruskal-Wallis identified more differentiating OTUs than did the other tests. The significant OTUs found by any method were combined into one feature set. As some OTUs were found significant by more than one method, just one of them was kept after combination. The collection of these common significant OTUs ("filtered" OTUs) was used for classification. The filtered OTUs for each of the two groups using UPARSE method are presented in tables below. Table 6 shows the 52 OTUs that resulted from all four statistical tests that were significantly different between the adenoma and cancer groups. Table 7 lists the 62 significantly detected OTUs between the healthy control and adenoma groups. Table 8 shows the 56 significant OTUs between the healthy control and cancer groups. Most of the significant OTUs for each of these binary comparisons belonged to the Firmicutes and Bacteroidetes phyla. As these two are the most abundant phyla of the gut microbiome, it is not unexpected that most of the significant OTUs are from these two phyla. There are some OTUs that have the same taxonomy but clustered as different OTUs which indicates that these are probably subtaxa (i.e., species or strains).

Table 6 The 52 significantly different OTUs between adenoma and cancer groups of Benchmark-1. Forty-Six OTUs are from Firmicutes and Bacteroidetes phyla. The most observed order is Clostridiales. Just a few OTUs are wholly classified up to species level. P: phylum; o: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU3	GG850870	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
2	OTU427	GG207615	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
3	OTU56	GG848088	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
4	OTU53	GG1105904	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
5	OTU300	GG207994	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
6	OTU1518	GG199710	p_Firmicutes;c_Clostridia;o_Clostridiales;
7	OTU884	GG343989	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
8	OTU291	GG211212	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
9	OTU341	GG974203	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;
10	OTU100	GG988932	p_Firmicutes;c_Clostridia;o_Clostridiales;
11	OTU1762	GG335523	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_ovatus
12	OTU788	GG298592	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;
13	OTU204	GG368448	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
14	OTU602	GG198839	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
15	OTU555	GG358112	p_Firmicutes;c_Clostridia;o_Clostridiales;
16	OTU240	GG358185	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
17	OTU165	GG308072	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
18	OTU356	GG796473	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;
19	OTU40	GG998587	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
20	OTU453	GG182797	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
21	OTU24	GG841108	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
22	OTU139	GG851797	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Clostridium;s_bolteae
23	OTU756	GG573053	p_Firmicutes;c_Clostridia;o_Clostridiales;
24	OTU1	GG1104433	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
25	OTU255	GG386273	p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Campylobacteraceae;g_Campylobacter;s_ureolyticus
26	OTU1365	GG4329459	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_fragilis
27	OTU713	GG512682	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
28	OTU1381	GG534926	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Anaerostipes;
29	OTU150	GG692756	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
30	OTU63	GG214651	p_Firmicutes;c_Clostridia;o_Clostridiales;
31	OTU392	GG215097	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
32	OTU752	GG167730	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Barnesiellaceae;
33	OTU231	GG259772	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;

No.	OTU Number	GreenGenes ID	Taxonomy
34	OTU723	GG338730	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
35	OTU205	GG365160	p_Firmicutes;c_Clostridia;o_Clostridiales;
36	OTU523	GG206523	p_Firmicutes;c_Clostridia;o_Clostridiales;
37	OTU1027	GG174893	p_Actinobacteria;c_Coriobacteriia;o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella;s_aerofaciens
38	OTU384	GG198720	p_Firmicutes;c_Clostridia;o_Clostridiales;
39	OTU95	GG363214	p_Firmicutes;c_Clostridia;o_Clostridiales;
40	OTU116	GG620319	p_Firmicutes;c_Clostridia;o_Clostridiales;
41	OTU197	GG864573	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_Lactobacillus;
42	OTU19	GG590945	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
43	OTU607	GG566233	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;
44	OTU298	GG114284	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
45	OTU229	GG362576	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
46	OTU1937	GG312882	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
47	OTU34	GG589710	p_Firmicutes;c_Clostridia;o_Clostridiales;
48	OTU482	GG193868	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
49	OTU219	GG535601	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
50	OTU551	GG358112	p_Firmicutes;c_Clostridia;o_Clostridiales;
51	OTU115	GG591825	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
52	OTU1179	GG302746	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;

Table 7 The 62 significantly different OTUs between adenoma and healthy control groups of Benchmark-1. Fifty-six OTUs are from the Firmicutes and Bacteroidetes phyla. The most observed order is Clostridiales. Just a few OTUs are entirely classified up to species level. P: phylum; o: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU10	GG581094	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
2	OTU113	GG367091	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
3	OTU1146	GG366735	p_Firmicutes;c_Clostridia;o_Clostridiales;
4	OTU115	GG591825	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
5	OTU1175	GG4004998	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
6	OTU1179	GG302746	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
7	OTU1194	GG555623	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_ovatus
8	OTU1233	GG838703	p_Firmicutes;c_Clostridia;o_Clostridiales;
9	OTU1275	GG178708	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
10	OTU1374	GG198044	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
11	OTU1392	GG839512	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_eggerthii
12	OTU1410	GG182469	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
13	OTU1422	GG196664	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_uniformis
14	OTU1478	GG276149	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides;
15	OTU1564	GG767952	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_Lactobacillus;s_zeae
16	OTU1888	GG180042	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
17	OTU189	GG358483	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
18	OTU20	GG369922	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
19	OTU2011	GG192963	p_Verrucomicrobia;c_Verrucomicrobiae;o_Verrucomicrobiales;f_Verrucomicrobiaceae;g_Akkermansia;s_muciniphila
20	OTU205	GG365160	p_Firmicutes;c_Clostridia;o_Clostridiales;
21	OTU224	GG4336947	p_Proteobacteria;c_Deltaproteobacteria;o_Desulfovibrionales;f_Desulfovibrionaceae;
22	OTU24	GG841108	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
23	OTU244	GG827743	p_Firmicutes;c_Clostridia;o_Clostridiales;
24	OTU249	GG299441	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
25	OTU28	GG470168	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea;
26	OTU290	GG848284	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
27	OTU318	GG536167	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
28	OTU346	GG519490	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
29	OTU368	GG349351	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
30	OTU38	GG316761	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_eggerthii
31	OTU380	GG157772	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;

No.	OTU Number	GreenGenes ID	Taxonomy
32	OTU384	GG198720	p_Firmicutes;c_Clostridia;o_Clostridiales;
33	OTU386	GG230405	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
34	OTU396	GG538344	p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomonadales;f_Pseudomonadaceae;g_Pseudomonas;s_veronii
35	OTU410	GG217109	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
36	OTU439	GG4410369	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
37	OTU44	GG198044	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
38	OTU447	GG2941399	p_Synergistetes;c_Synergistia;o_Synergistales;f_Synergistaceae;g_Synergistes;
39	OTU46	GG846409	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
40	OTU464	GG342105	p_Firmicutes;c_Clostridia;o_Clostridiales;
41	OTU478	GG290465	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
42	OTU481	GG360653	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_ovatus
43	OTU5	GG470690	p_Euryarchaeota;c_Methanobacteria;o_Methanobacteriales;f_Methanobacteriaceae;g_Methanobrevibacter;
44	OTU520	GG174885	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
45	OTU536	GG550814	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
46	OTU551	GG358112	p_Firmicutes;c_Clostridia;o_Clostridiales;
47	OTU555	GG358112	p_Firmicutes;c_Clostridia;o_Clostridiales;
48	OTU577	GG185927	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
49	OTU591	GG349876	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
50	OTU67	GG291420	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Eubacterium;s_biforme
51	OTU709	GG273967	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
52	OTU711	GG276158	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
53	OTU713	GG512682	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
54	OTU714	GG246717	p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Campylobacteraceae;g_Campylobacter;
55	OTU721	GG182512	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
56	OTU760	GG178238	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
57	OTU788	GG298592	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;
58	OTU818	GG266621	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
59	OTU942	GG444962	p_Firmicutes;c_Bacilli;o_Bacillales;f_Planococcaceae;
60	OTU953	GG320156	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea;
61	OTU96	GG591635	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
62	OTU99	GG523919	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;

Table 8 The 56 significantly different OTUs between healthy control and cancer groups of Benchmark-1. Fifty-Two OTUs are from the Firmicutes and Bacteroidetes phyla. The most observed order is Clostridiales. Just a few OTUs are wholly classified up to species level. P: phylum; o: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU1094	GG323818	p_Firmicutes;c_Clostridia;o_Clostridiales;
2	OTU1141	GG364582	p_Firmicutes;c_Clostridia;o_Clostridiales;
3	OTU1146	GG366735	p_Firmicutes;c_Clostridia;o_Clostridiales;
4	OTU1147	GG546876	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
5	OTU123	GG848492	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
6	OTU139	GG851797	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Clostridium;s_bolteae
7	OTU1392	GG839512	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_eggerthii
8	OTU1533	GG364334	p_Verrucomicrobia;c_Verrucomicrobiae;o_Verrucomicrobiales;f_Verrucomicrobiaceae;g_Akkermansia;s_muciniphila
9	OTU17	GG579112	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_caccae
10	OTU1730	GG581554	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
11	OTU174	GG336761	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
12	OTU189	GG358483	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
13	OTU19	GG590945	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
14	OTU219	GG535601	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
15	OTU2195	GG4225004	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea;
16	OTU224	GG4336947	p_Proteobacteria;c_Deltaproteobacteria;o_Desulfovibrionales;f_Desulfovibrionaceae;
17	OTU229	GG362576	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
18	OTU2368	GG183852	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
19	OTU244	GG827743	p_Firmicutes;c_Clostridia;o_Clostridiales;
20	OTU2497	GG1105984	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
21	OTU2517	GG484304	p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium;
22	OTU2524	GG179905	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
23	OTU253	GG293869	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea;
24	OTU263	GG33987	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
25	OTU29	GG172962	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;
26	OTU291	GG211212	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
27	OTU298	GG114284	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
28	OTU300	GG207994	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
29	OTU318	GG536167	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
30	OTU340	GG350666	p_Firmicutes;c_Clostridia;o_Clostridiales;
31	OTU37	GG538796	p_Firmicutes;c_Clostridia;o_Clostridiales;
32	OTU373	GG175654	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
33	OTU38	GG316761	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_eggerthii

No.	OTU Number	GreenGenes ID	Taxonomy
34	OTU40	GG998587	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
35	OTU419	GG4341119	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
36	OTU421	GG313524	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
37	OTU427	GG207615	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
38	OTU474	GG594304	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
39	OTU536	GG550814	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
40	OTU56	GG848088	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
41	OTU57	GG368969	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;
42	OTU577	GG185927	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
43	OTU602	GG198839	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
44	OTU65	GG586680	p_Firmicutes;c_Clostridia;o_Clostridiales;
45	OTU67	GG291420	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Eubacterium;s_biforme
46	OTU714	GG246717	p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Campylobacteraceae;g_Campylobacter;
47	OTU721	GG182512	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
48	OTU752	GG167730	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Barnesiellaceae;
49	OTU768	GG2963287	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
50	OTU788	GG298592	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;
51	OTU85	GG368350	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;s_producta
52	OTU866	GG177911	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;
53	OTU913	GG291518	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
54	OTU95	GG363214	p_Firmicutes;c_Clostridia;o_Clostridiales;
55	OTU961	GG365456	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
56	OTU982	GG192741	p_Firmicutes;c_Clostridia;o_Clostridiales;

Some OTUs were commonly detected in all three comparisons. Among these OTUs were the family of Fusobacteriaceae, the order of Clostridiales, and the genus of *Blautia*. One could hypothesize that these OTUs may stimulate the malignant progression of the disease, but this concept would have to be validated by functional experiments. In samples collected from CRC patients, some of these OTUs were increased in their abundance, and some were decreased. The bar plots of the normalized abundance of each binary group of Benchmark-1 are depicted in Figure 25, Figure 26, and Figure 27.

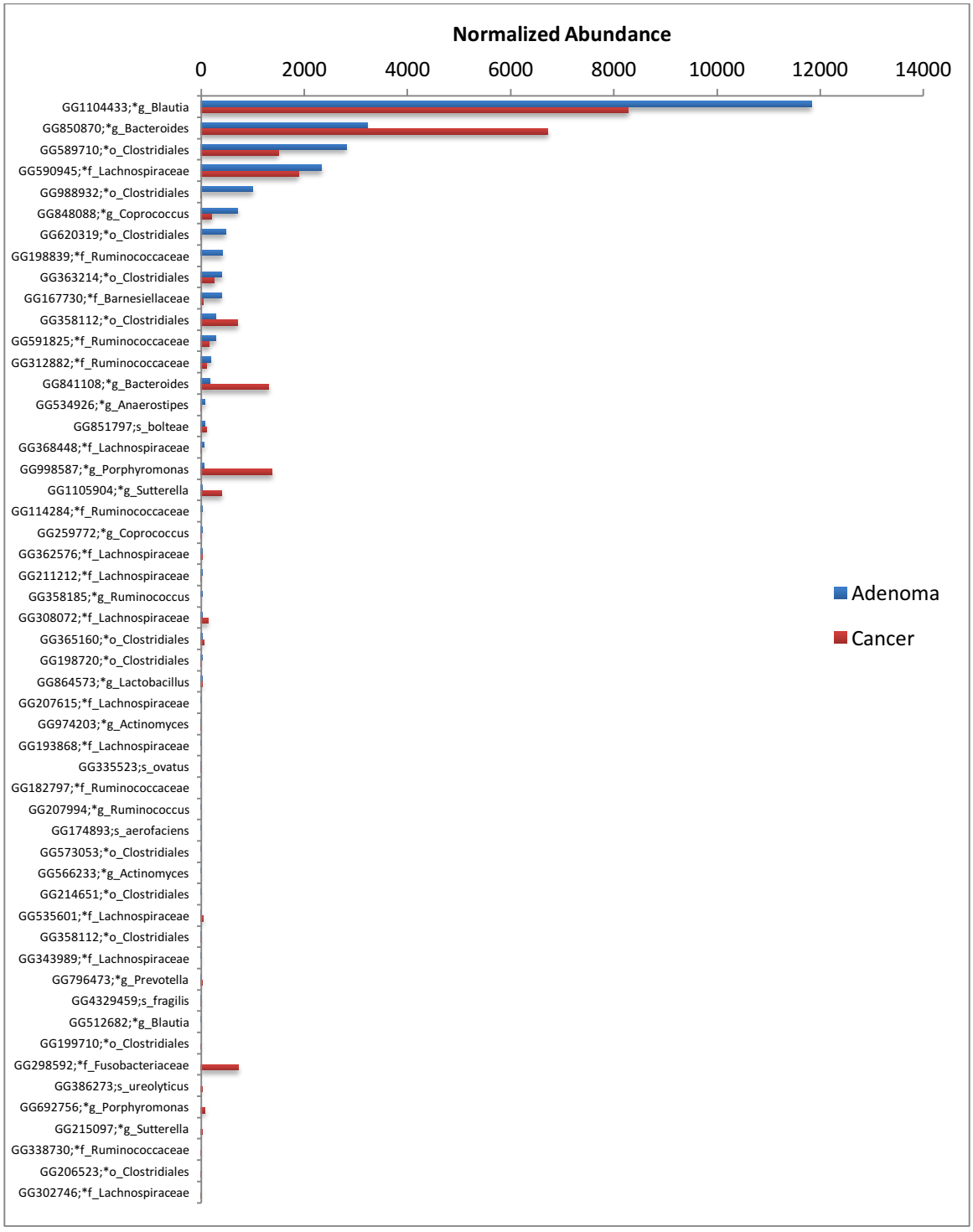


Figure 25 Change of significantly different OTUs between adenoma and cancer groups in Benchmark-1. The abundance of some bacterial taxa decreased, and some increased at cancer state compared to adenoma. O: order; f: family; g: genus; s: species.

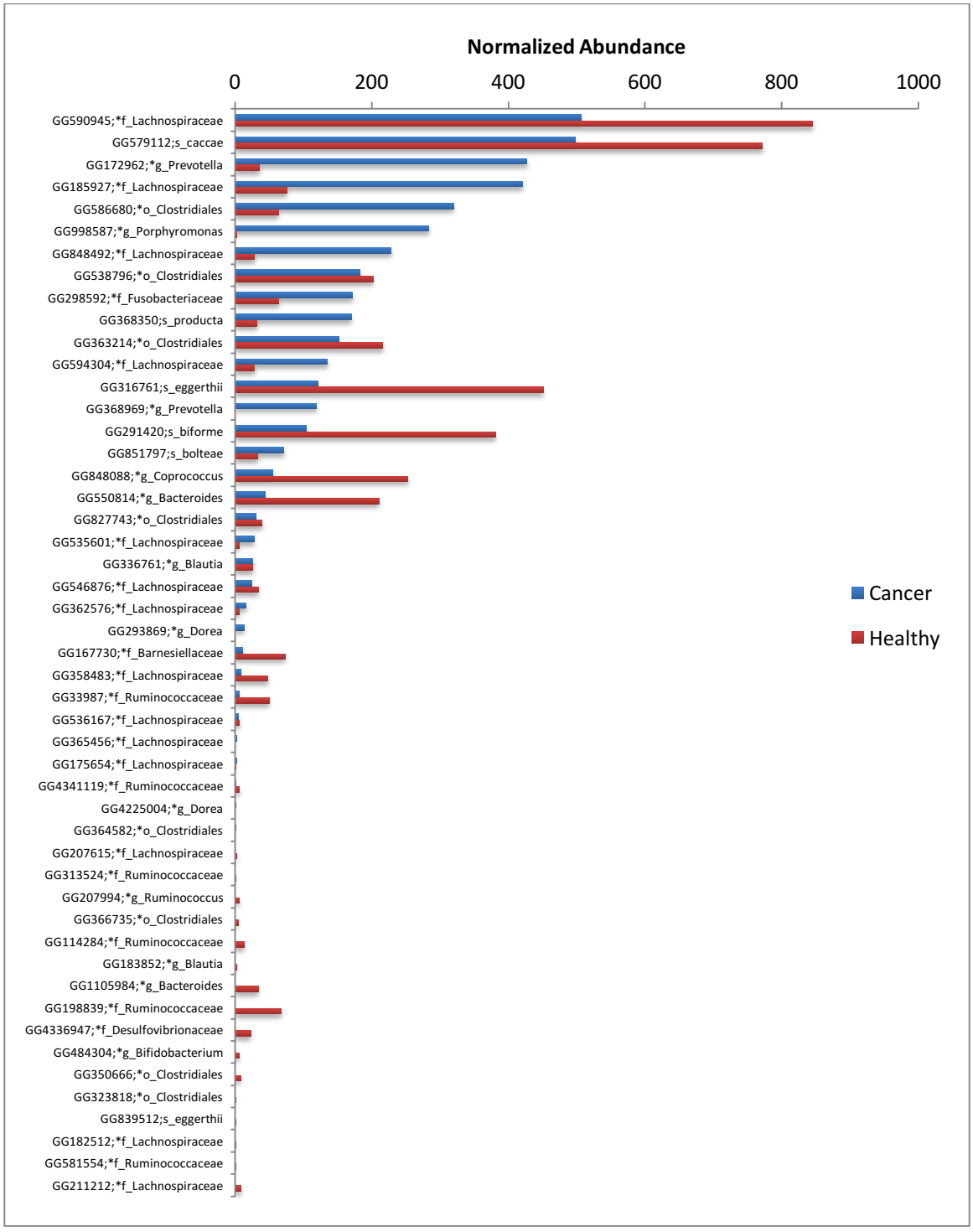


Figure 26 Change of significantly different OTUs between healthy control and cancer groups in Benchmark-1. The abundance of some bacterial taxa decreased, and some increased at cancer state compared to healthy control. O: order; f: family; g: genus; s: species.

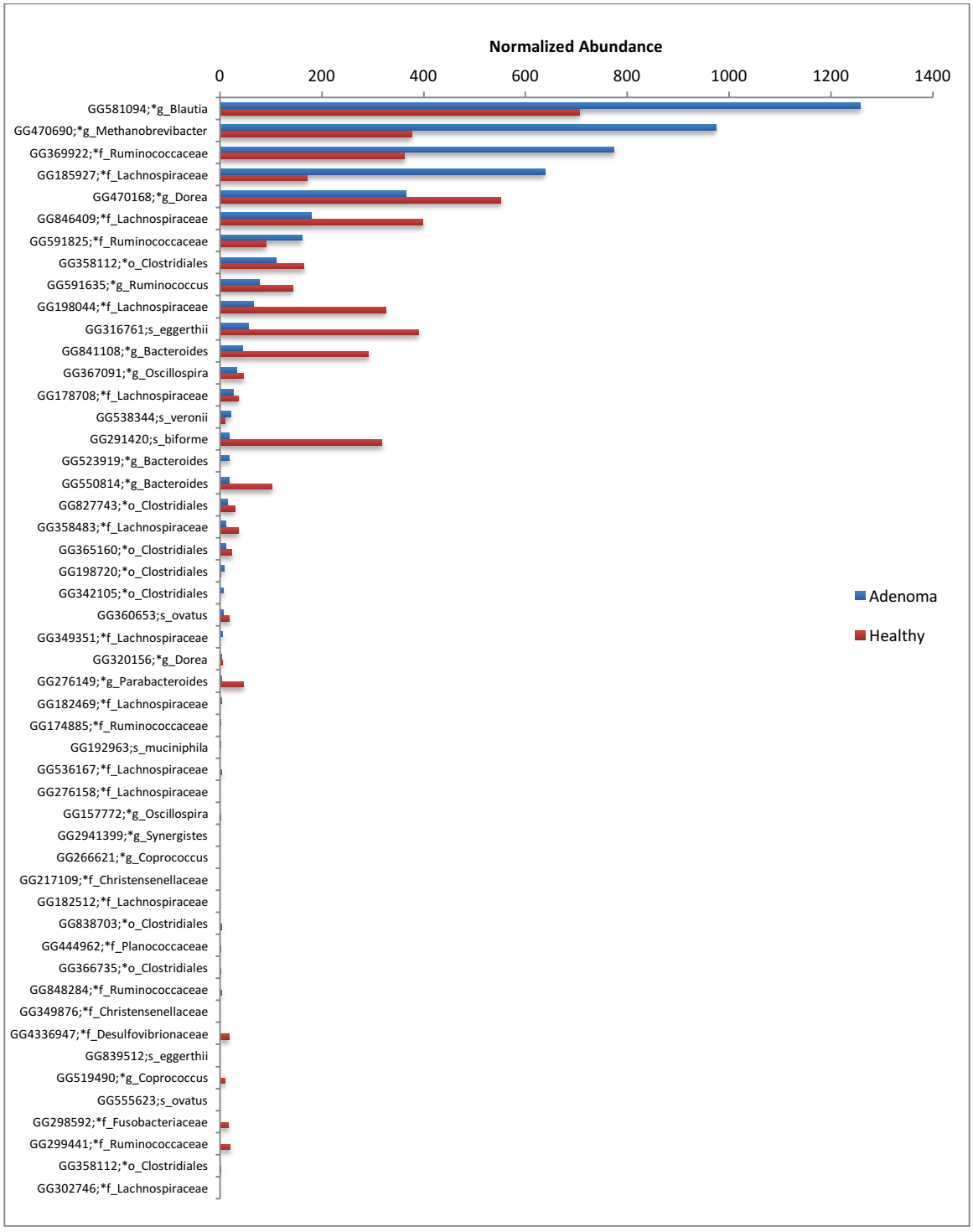


Figure 27 Change of significantly different OTUs between adenoma and healthy control groups in Benchmark-1.

The abundance of some bacterial taxa decreased, and some increased at adenoma state compared to healthy control. O: order; f: family; g: genus; s: species.

4.1.5 Benchmark-1 classification

The classification was performed using the OTU abundance table of each OTU selection method. The classification was performed once with the all of the OTUs as classification features (referred to as the "raw feature set"), and once with significant OTUs found by statistical tests as classification features (referred to as the "filtered feature set"). Several classification methods were employed including Naïve Bayes, Random forest, K Nearest Neighbor (KNN), Classification tree, Logistic regression, and Neural Network.

4.1.6 Benchmark-1 classification validation

After applying different classifiers, classification validation was performed with the 10-fold cross-validation method. At the next step, trained classifiers were evaluated on test datasets. Results from the UPARSE method are shown in Table 9. As the classification results on UPARSE OTUs showed better specificity, sensitivity, and accuracy compared to other methods, we will just describe the results of UPARSE here.

The classification accuracy (CA), sensitivity (Sens), specificity (Spec), and area under the ROC curve (AUC) were different for each classification method and is summarized in Table 9. The first three raw tables are the validation results of classifiers that are made by using all OTUs as features (raw feature set). The second set of tables show validation results when the significant OTUs used as features (filtered feature set). The last tables show the performance of the classifiers produced by significant features on the three different test dataset each of which is 20% of the dataset.

Table 9 The UPARSE-guided validation of the Benchmark-1 classification.
The top tables show the validation results of the classifiers produced using all OTUs. The middle tables are the results of just significant OTUs as classification features. Bottom tables are the results of applying the classifiers produced by significant OTUs on test datasets. Classifiers' performance improved for all binary groups when significant OTUs were used as classification features instead of all OTUs. CA: classification accuracy; Sens: sensitivity; Spec: specificity; AUC: area under the curve, KNN: K nearest neighbor.

Benchmark-1_UPARSE

Raw_Adenoma_Cancer				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.60	0.87	0.33	0.57
Random Forest	0.60	0.57	0.63	0.62
kNN	0.58	0.52	0.63	0.61
Classification Tree	0.68	0.70	0.67	0.68
Logistic regression	0.62	0.70	0.54	0.72
Neural Network	0.64	0.70	0.58	0.73
SVM	0.47	0.26	0.67	0.50

Raw_Healthy_Adenoma				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.68	0.46	0.91	0.68
Random Forest	0.60	0.67	0.52	0.67
kNN	0.54	0.92	0.13	0.66
Classification Tree	0.58	0.58	0.57	0.58
Logistic regression	0.64	0.71	0.57	0.62
Neural Network	0.64	0.71	0.57	0.62
SVM	0.51	0.75	0.26	0.50

Raw_Healthy_Cancer				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.51	0.48	0.52	0.53
Random Forest	0.48	0.52	0.43	0.48
kNN	0.56	0.83	0.30	0.56
Classification Tree	0.50	0.48	0.52	0.49
Logistic regression	0.60	0.74	0.48	0.65
Neural Network	0.61	0.78	0.43	0.66
SVM	0.48	0.22	0.74	0.50

Filtered_Adenoma_Cancer				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.90	0.87	0.92	0.98
Random Forest	0.89	0.87	0.92	0.97
kNN	0.85	0.83	0.88	0.91
Classification Tree	0.84	0.96	0.71	0.84
Logistic regression	0.89	0.87	0.92	0.94
Neural Network	0.89	0.87	0.92	0.95
SVM	0.62	0.39	0.83	0.72

Filtered_Healthy_Adenoma				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.90	0.88	0.91	0.95
Random Forest	0.70	0.83	0.57	0.81
kNN	0.87	1.00	0.74	0.96
Classification Tree	0.62	0.67	0.57	0.64
Logistic regression	0.94	0.96	0.91	0.99
Neural Network	0.96	1.00	0.91	1.00
SVM	0.74	0.50	1.00	0.87

Filtered_Healthy_Cancer				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.87	0.91	0.83	0.93
Random Forest	0.76	0.83	0.70	0.89
kNN	0.80	0.83	0.78	0.84
Classification Tree	0.61	0.61	0.61	0.62
Logistic regression	0.87	0.83	0.91	0.94
Neural Network	0.91	0.87	0.96	0.95
SVM	0.37	0.43	0.30	0.35

On test data				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	1	1	1	1
Random Forest	0.83	0.66	1	0.94
kNN	1	1	1	1
Classification Tree	0.83	0.66	1	0.83
Logistic regression	1	1	1	1
Neural Network	1	1	1	1
SVM	0.75	0.5	1	1

On test data				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.91	1	0.83	0.83
Random Forest	0.66	1	0.33	0.97
kNN	0.83	0.83	0.83	0.88
Classification Tree	0.33	0.5	0.16	0.33
Logistic regression	0.91	0.83	1	1
Neural Network	0.91	0.83	1	1
SVM	0.75	0.5	1	0.88

On test data				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.91	0.83	1	0.94
Random Forest	0.83	1	0.66	1
kNN	0.75	0.83	0.66	0.72
Classification Tree	0.66	0.83	0.5	0.65
Logistic regression	0.75	1	0.5	0.88
Neural Network	0.75	1	0.5	0.97
SVM	0.58	0.33	0.83	0.58

The classifiers with reasonably good to excellent CA, sensitivity, and specificity are the Naïve Bayes, the neural network, the K-nearest neighbor, the logistic regression, and the random forest.

ROC curves of Benchmark-1 study are shown in Figure 28, Figure 29, and Figure 30. The curves of the five above classifiers with higher areas under the curve are shown. For this dataset, the classifiers with the highest area (higher than 80%) were Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). Neural network and Naïve Bayes were the best performings in this dataset.

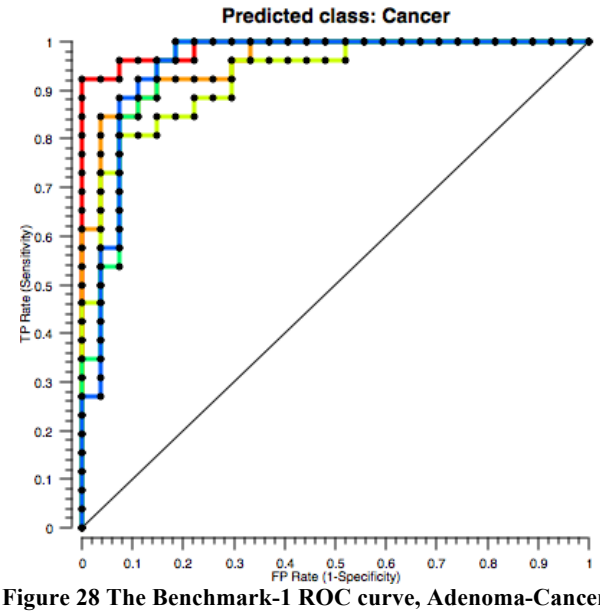


Figure 28 The Benchmark-1 ROC curve, Adenoma-Cancer.
The straight line represents the null model. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The x-axis is the true positive rate, and the y-axis is the false positive rate. Naïve Bayes is the best performing classifier in this binary group with AUC= 0.98.

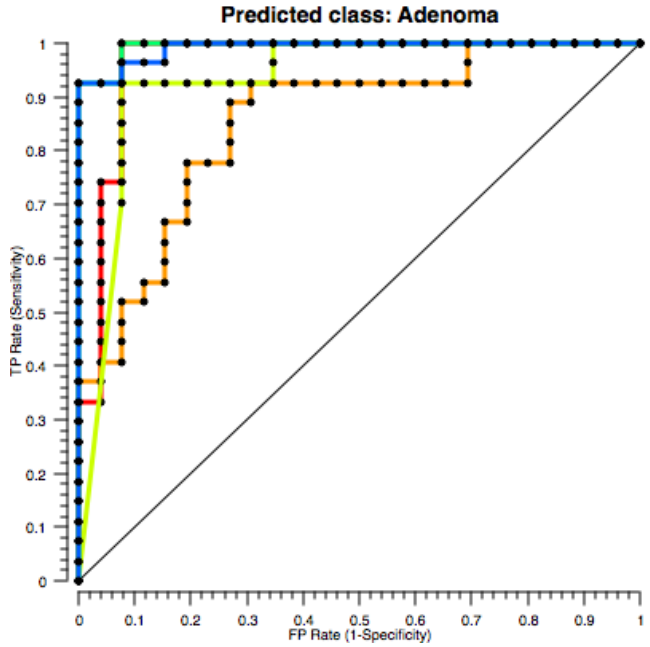


Figure 29 The Benchmark-1 ROC curve, Healthy-Adenoma.
The straight line represents the null model. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The x-axis is the true positive rate, and the y-axis is the false positive rate. The neural network is the best performing classifier in this binary group with AUC= 1.

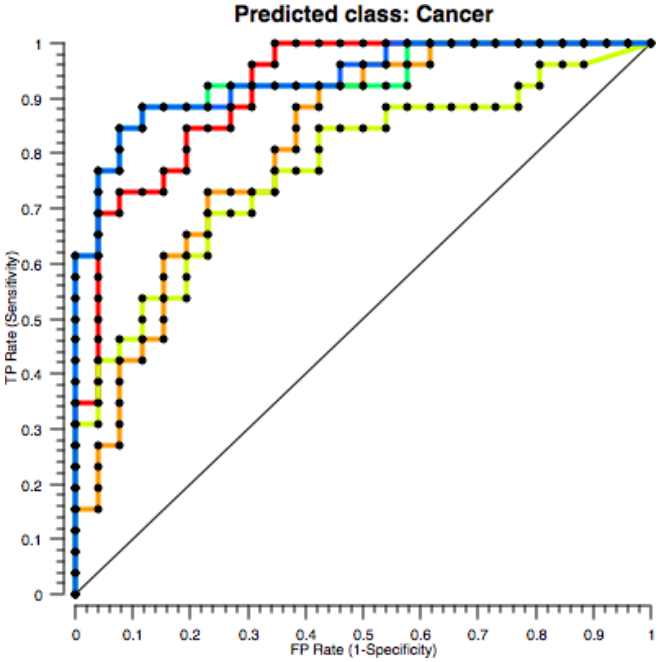


Figure 30 The Benchmark-1 ROC curve, Healthy-Cancer.
The straight line represents the null model. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The x-axis is the true positive rate, and the y-axis is the false positive rate. The neural network is the best performing classifier in this binary group with AUC= 0.95.

4.2 Benchmark-2 results

The sequences from Wu et al. 2013 study were preprocessed with the same parameters and three OTU selection methods as Benchmark-1. Table 10 shows that the number of original reads for this benchmark was 727,860. After preprocessing, removing noises, removing undesirable lineages, and low-quality reads, 523,931 reads remained for the UPARSE and UPGMA methods and 709,713 reads remained for the UCLUST method. The number of detected OTUs were 1208, 6,975, and 33,533 OTUs for UPARSE, UPGMA, and UCLUST, respectively. As in Benchmark-1, UPARSE gave us the lowest number of centroids and UCLUST returned the highest number of centroids.

Table 10 Comparing three OTU selection methods using Benchmark-2.
The UPARSE returned the lowest number of OTUs while UCLUST found the highest.

	UPARSE	UPGMA	UCLUST
No. of original reads (for alignment)	727,860	727,860	727,860
No. of reads for clustering	523,931	523,931	709,713
No. of detected OTUs	1208	6975	33533

4.2.1 Benchmark-2 rarefaction

As in Benchmark 1, rarefaction curves were prepared to check the sequencing depth which is shown in Figure 31. For both healthy control and CRC groups, rarefaction plots showed good sequencing depth as the curves were almost parallel to the x-axis. Therefore,

we can be confident that the depth of coverage was sufficient to identify the true number of OTUs in the samples.

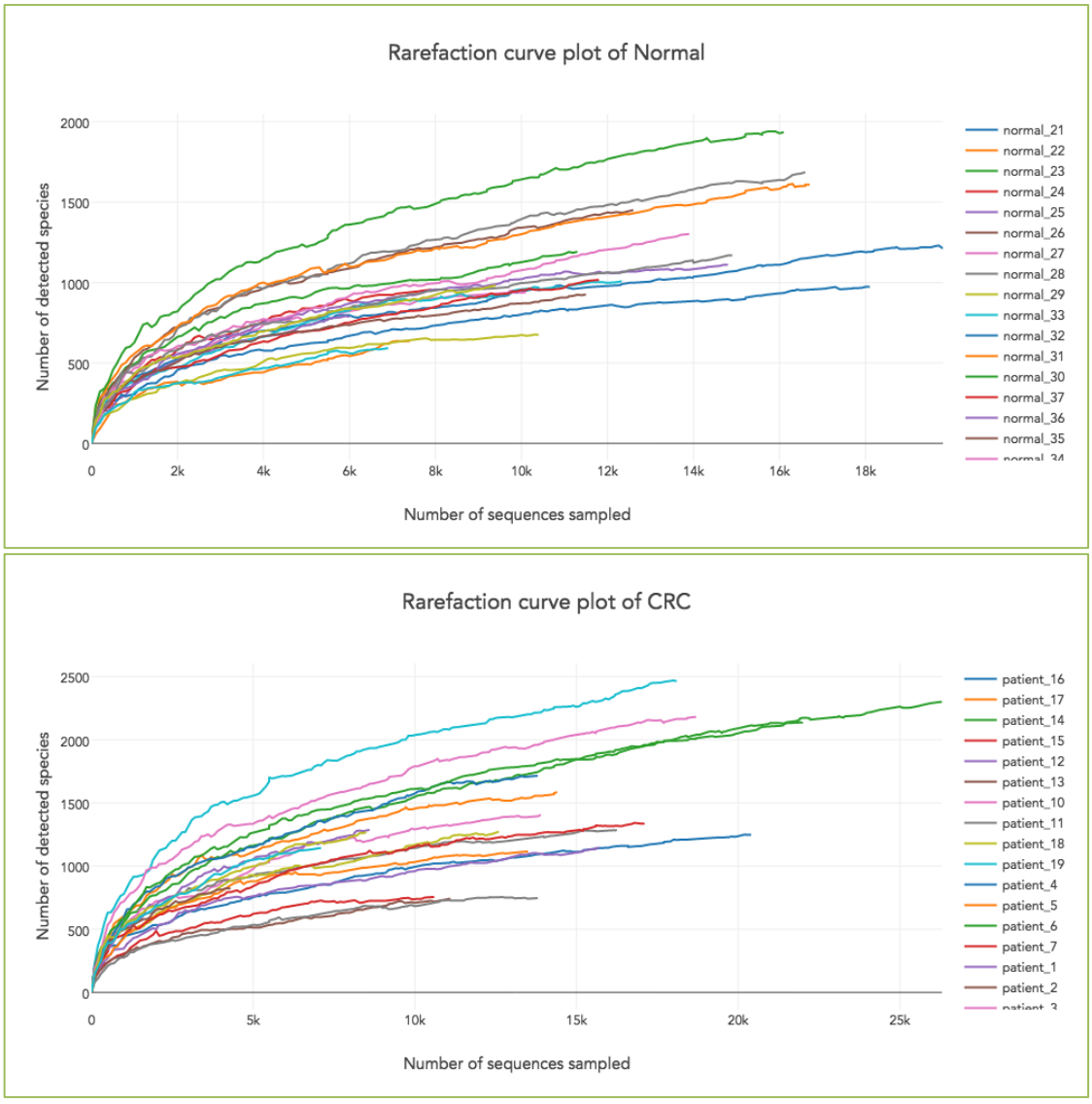


Figure 31 The rarefaction plots of Benchmark-2.
The plots have reached a plateau then the sequencing depth is enough.

4.2.2 Benchmark-2 alpha diversity

The average alpha diversity indices for each group of healthy control and CRC subjects were summarized in Table 11. The pattern of alpha diversity is the same for all methods although the absolute values are different. In all three methods, the diversity of the CRC group is higher than the healthy control group.

Table 11 Benchmark-2 alpha diversity.
The pattern of alpha diversity is the same for all three methods although the absolute values are different. The diversity of the CRC group is higher than the healthy control group.

	UPARSE		**UPGMA**		**UCLUST**	
Diversity index	Average in group		Average in group		Average in group	
Group	CRC	Healthy	CRC	Healthy	CRC	Healthy
nseq	19206	16504	19514	16754	19595	16797
Shannon	3.5	3	4.09	3.68	4.89	4.53
Simpson	0.07	0.13	0.04	0.06	0.02	0.04
invSimpson	17.69	10.94	33.30	20.68	55.62	37.97
Sobs	228	203	301	250	950	743

4.2.3 Benchmark-2 beta diversity

As in Benchmark 1, phylogenetic Unifrac and nonphylogenetic Bray-Curtis metrics were used to analyze beta diversity. The PCoA visualization plots indicated that UniFrac

was not successful in discriminating cancer and healthy control groups as shown in Figure 32.

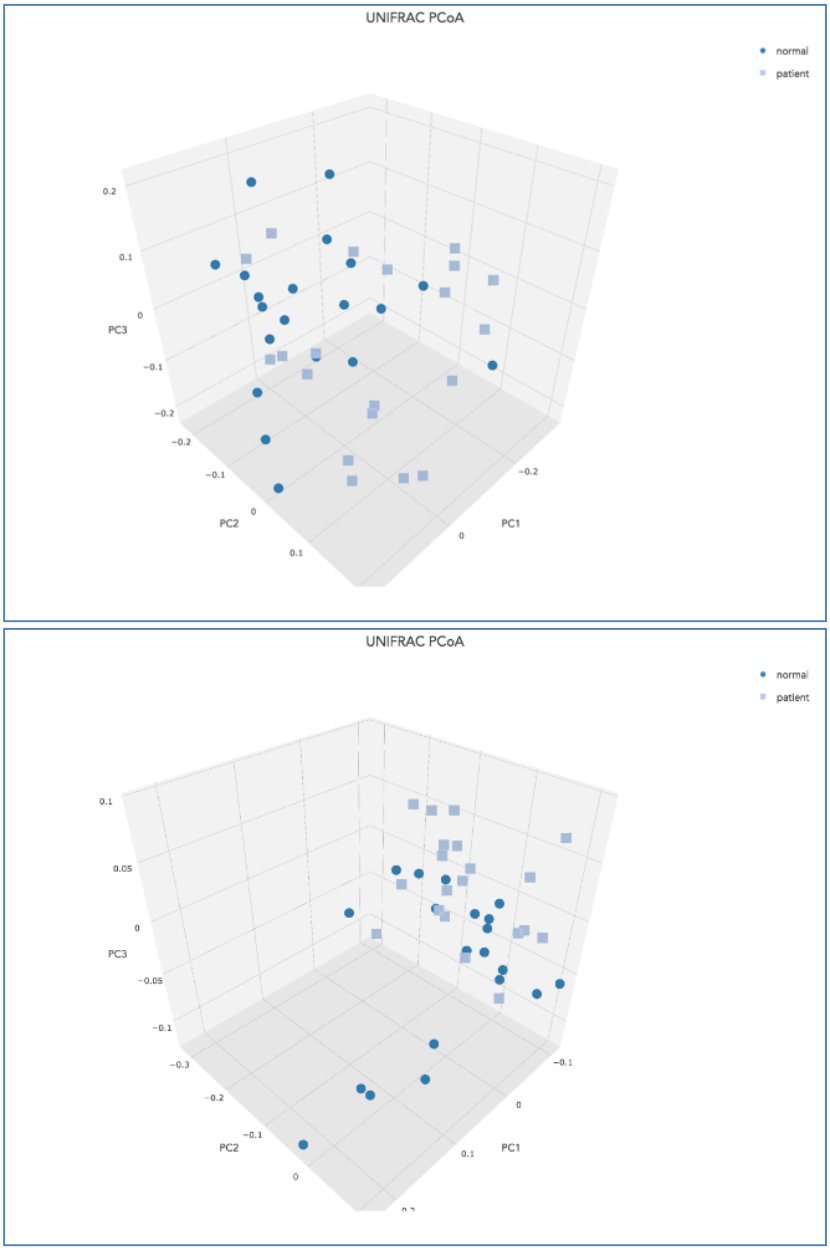


Figure 32 The Benchmark-2 UniFrac PCoA plots.
Neither unweighted (top) nor weighted (bottom) UniFrac were successful in discriminating healthy control and cancer groups. The axes are principal components.

Similarly, the PCoA visualization in Figure 33 reveals that Bray-Curtis metric did not produce significant clustering between the healthy control and the CRC groups either.

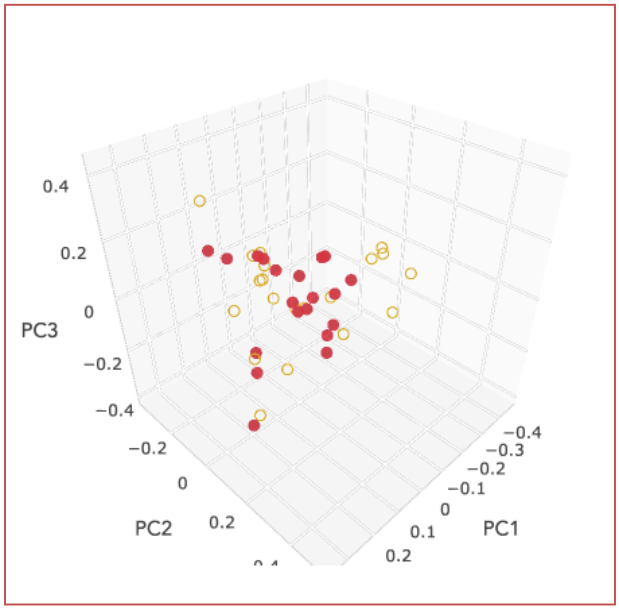


Figure 33 The Benchmark-2 Bray-Curtis PCoA plot.
There is no clear separation between CRC and healthy control groups using the first three principal components. The axes are principal components.

4.2.4 Benchmark-2 significant OTUs

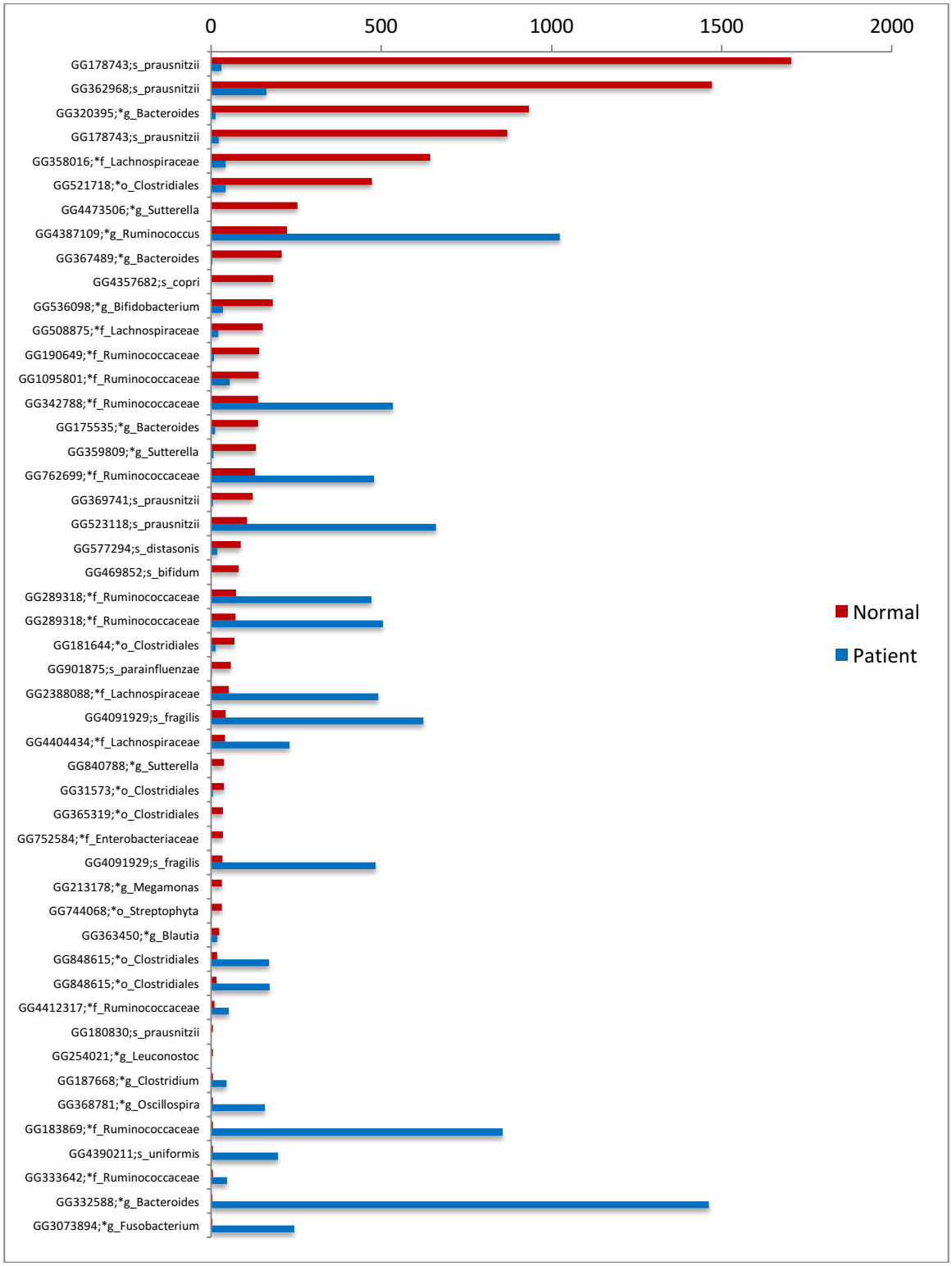
As in Benchmark-1, Kruskal-Wallis, MetaStats, LEfSe, and Indicator methods were applied on UPARSE OTU table to find significant OTUs. Significant OTUs detected by each of these methods were combined and represented in Table 12. In summary, 94 significant OTUs were detected, and the vast majority of them were from Firmicutes and Bacteroidetes phyla. Only 16 OTUs were identified at species level. Figure 34 shows the direction of the bacterial change in the cancer group compared to the healthy control group. Some of the significant taxa increased in the disease state and some decreased.

Table 12 The significantly different OTUs between healthy control and cancer groups of Benchmark-2. From Ninety-Four significant OTUs that have been detected, Eighty-Four OTUs are from Firmicutes and Bacteroidetes phyla. There are some OTUs that have the same taxonomy but categorized in different OTU groups which means they are subtaxa. O: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU1	GG332588	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
2	OTU1020	GG320395	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
3	OTU1029	GG181644	p_Firmicutes;c_Clostridia;o_Clostridiales;
4	OTU1032	GG298535	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
5	OTU1058	GG901875	p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Haemophilus;s_parainfluenzae
6	OTU11	GG4091929	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_fragilis
7	OTU1113	GG254021	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Leuconostocaceae;g_Leuconostoc;
8	OTU1118	GG4363660	p_Firmicutes;c_Clostridia;o_Clostridiales;
9	OTU1132	GG533847	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
10	OTU1135	GG3667016	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_uniformis
11	OTU1143	GG928652	p_Firmicutes;c_Bacilli;o_Gemellales;f_Gemellaceae;g_Gemella;
12	OTU1163	GG175535	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
13	OTU12	GG1055711	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
14	OTU120	GG521718	p_Firmicutes;c_Clostridia;o_Clostridiales;
15	OTU128	GG4309698	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;g_Fusobacterium;
16	OTU136	GG801964	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister;
17	OTU139	GG289318	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
18	OTU14	GG1055711	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
19	OTU142	GG358016	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
20	OTU154	GG178743	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
21	OTU156	GG523357	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;s_copri
22	OTU161	GG692756	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
23	OTU163	GG1896087	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;g_Fusobacterium;
24	OTU167	GG342788	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
25	OTU18	GG4091929	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_fragilis
26	OTU180	GG621651	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;
27	OTU183	GG762699	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
28	OTU187	GG216599	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
29	OTU196	GG692756	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas;
30	OTU202	GG848615	p_Firmicutes;c_Clostridia;o_Clostridiales;

No.	OTU Number	GreenGenes ID	Taxonomy
31	OTU208		
32	OTU225	GG621651	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;
33	OTU232	GG422136	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;s_intermedia
34	OTU247	GG536098	p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium;
35	OTU264	GG848615	p_Firmicutes;c_Clostridia;o_Clostridiales;
36	OTU266	GG359809	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
37	OTU272	GG368781	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
38	OTU278	GG183869	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
39	OTU284	GG1095801	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
40	OTU294	GG4318033	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_Granulicatella;
41	OTU295	GG333642	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
42	OTU299	GG4357682	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;s_copri
43	OTU315	GG368781	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
44	OTU333	GG178743	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
45	OTU335	GG15712	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Clostridium;s_ramosum
46	OTU336	GG187668	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_Clostridium;
47	OTU342	GG190649	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
48	OTU347	GG4455436	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister;
49	OTU360	GG752584	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
50	OTU363	GG849362	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Odoribacteraceae;g_Butyricimonas;
51	OTU364	GG744068	p_Cyanobacteria;c_Chloroplast;o_Streptophyta;
52	OTU365	GG918720	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Parvimonas;
53	OTU373		
54	OTU386	GG4478840	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
55	OTU394	GG362968	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
56	OTU424	GG2855173	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
57	OTU427	GG953126	p_Firmicutes;c_Bacilli;o_Gemellales;f_Gemellaceae;
58	OTU445	GG508875	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
59	OTU452	GG4412317	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
60	OTU471	GG367215	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Holdemania;
61	OTU478	GG540055	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
62	OTU514	GG574200	p_Proteobacteria;c_Betaproteobacteria;o_Neisseriales;f_Neisseriaceae;g_Eikenella;
63	OTU517	GG189338	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;
64	OTU520	GG840788	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
65	OTU524	GG469852	p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium;s_bifidum

No.	OTU Number	GreenGenes ID	Taxonomy
66	OTU534	GG446135	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
67	OTU558	GG516733	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
68	OTU578	GG4387109	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;
69	OTU592	GG582284	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
70	OTU602	GG533579	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides;
71	OTU614	GG850329	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_caccae
72	OTU63	GG523118	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
73	OTU640		
74	OTU670	GG3426658	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
75	OTU675	GG213178	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Megamonas;
76	OTU696	GG369741	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
77	OTU700	GG198814	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
78	OTU720	GG583117	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
79	OTU737	GG3073894	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;g_Fusobacterium;
80	OTU749	GG366987	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
81	OTU756	GG4404434	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
82	OTU769	GG180830	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
83	OTU78	GG2388088	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
84	OTU793	GG846373	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
85	OTU816	GG365319	p_Firmicutes;c_Clostridia;o_Clostridiales;
86	OTU826	GG363450	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
87	OTU836	GG577294	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides;s_distasonis
88	OTU853	GG367489	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
89	OTU888	GG357046	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
90	OTU90	GG289318	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
91	OTU903	GG31573	p_Firmicutes;c_Clostridia;o_Clostridiales;
92	OTU91	GG4473506	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
93	OTU920	GG368776	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_caccae
94	OTU928	GG4390211	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_uniformis


0
500
1000
1500
2000
GG178743;s_prausnitzii
GG362968;s_prausnitzii
GG320395;*g_Bacteroides
GG178743;s_prausnitzii
GG358016;*f_Lachnospiraceae
GG521718;*o_Clostridiales
GG4473506;*g_Sutterella
GG4387109;*g_Ruminococcus
GG367489;*g_Bacteroides
GG4357682;s_copri
GG536098;*g_Bifidobacterium
GG508875;*f_Lachnospiraceae
GG190649;*f_Ruminococcaceae
GG1095801;*f_Ruminococcaceae
GG342788;*f_Ruminococcaceae
GG175535;*g_Bacteroides
GG359809;*g_Sutterella
GG762699;*f_Ruminococcaceae
GG369741;s_prausnitzii
GG523118;s_prausnitzii
GG577294;s_distasonis
GG469852;s_bifidum
GG289318;*f_Ruminococcaceae
GG289318;*f_Ruminococcaceae
GG181644;*o_Clostridiales
GG901875;s_parainfluenzae
GG2388088;*f_Lachnospiraceae
GG4091929;s_fragilis
GG4404434;*f_Lachnospiraceae
GG840788;*g_Sutterella
GG31573;*o_Clostridiales
GG365319;*o_Clostridiales
GG752584;*f_Enterobacteriaceae
GG4091929;s_fragilis
GG213178;*g_Megamonas
GG744068;*o_Streptophyta
GG363450;*g_Blautia
GG848615;*o_Clostridiales
GG848615;*o_Clostridiales
GG4412317;*f_Ruminococcaceae
GG180830;s_prausnitzii
GG254021;*g_Leuconostoc
GG187668;*g_Clostridium
GG368781;*g_Oscillospira
GG183869;*f_Ruminococcaceae
GG4390211;s_uniformis
GG333642;*f_Ruminococcaceae
GG332588;*g_Bacteroides
GG3073894;*g_Fusobacterium
Normal
Patient

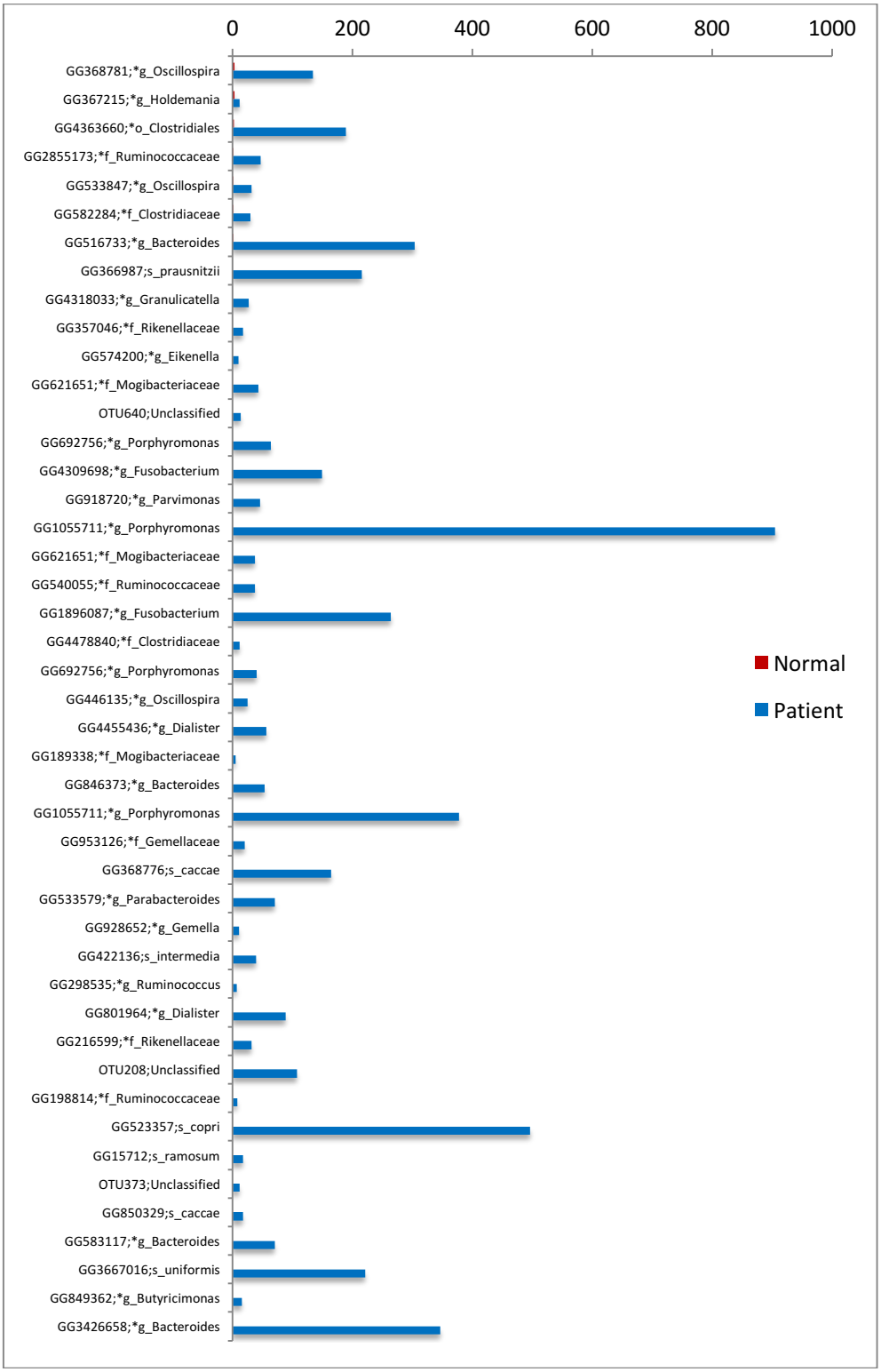


Figure 34 Change of significant OTUs between healthy control and cancer groups in Benchmark-2. Some taxa increased, and some decreased in the cancer state compared to the healthy control. O: order; f: family; g: genus; s: species.

4.2.5 Benchmark-2 classification

The classification was performed using the OTU abundance table produced by each of the three OTU selection methods, and different methods of classification were applied to the data as described for Benchmark-1, the classification was performed with all of the OTUs as classification features (raw feature set) and with significant OTUs detected by Kruskal-Wallis, MetaStats, LEfSe, and Indicator (filtered feature set).

4.2.6 Benchmark-2 classification validation

After utilizing the different classifiers, classification validation was executed with a 10-fold cross-validation method and then classifiers were validated on the test datasets. UPARSE results were better than UPGMA and UCLUST with respect to the sensitivity, specificity, and classification accuracy and only results from the UPARSE method are shown in Table 13. Classifier performance improved when we used significant OTUs to build classifiers. Among these classification methods, Naïve Bayes has the highest accuracy, sensitivity, and specificity for this dataset.

Table 13 The Benchmark-2 classification validation results.
The top table shows validation results of the classifiers produced using all OTUs. The middle table used just significant OTUs as classification features. The bottom is the results of applying the classifiers made by significant OTUs on the test dataset. Classifiers' performance improved when significant OTUs used as classification features instead of all OTUs. CA: classification accuracy; Sens: sensitivity; Spec: specificity; AUC: area under curve.

Benchmark-2_UPARSE

Raw_Healthy_Cancer				
Classification method	**CA**	**Sens**	**Spec**	**AUC**
Naïve Bayes	0.75	0.69	0.81	0.78
Random Forest	0.79	0.69	0.88	0.90
kNN	0.62	0.38	0.88	0.73
Classification Tree	0.50	0.44	0.56	0.55
Logistic regression	0.68	0.63	0.75	0.70
Neural Network	0.68	0.63	0.75	0.65
SVM	0.37	0.25	0.50	0.50

Filtered_Healthy_Cancer				
Classification method	**CA**	**Sens**	**Spec**	**AUC**
Naïve Bayes	0.97	1.00	0.94	1.00
Random Forest	0.91	0.88	0.94	0.95
kNN	0.86	0.81	0.88	1.00
Classification Tree	0.66	0.69	0.63	0.63
Logistic regression	0.94	1.00	0.88	1.00
Neural Network	0.94	1.00	0.88	1.00
SVM	0.43	0.25	0.63	0.40

On test data				
Classification method	**CA**	**Sens**	**Spec**	**AUC**
Naïve Bayes	1	1	1	1
Random Forest	0.875	0.75	1	1
kNN	0.875	0.75	1	1
Classification Tree	0.625	0.5	0.75	0.625
Logistic regression	1	1	1	1
Neural Network	0.875	0.75	1	1
SVM	0.5	0	1	0.25

The ROC curve from the UPARSE significant OTUs is shown in Figure 35. For this dataset, five classifiers with the highest area under the curve were the Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural

Network (blue). Based on the ROC curve, the Naïve Bayes, logistic regression, and neural network are the best performing classifier in this dataset with AUC=1.

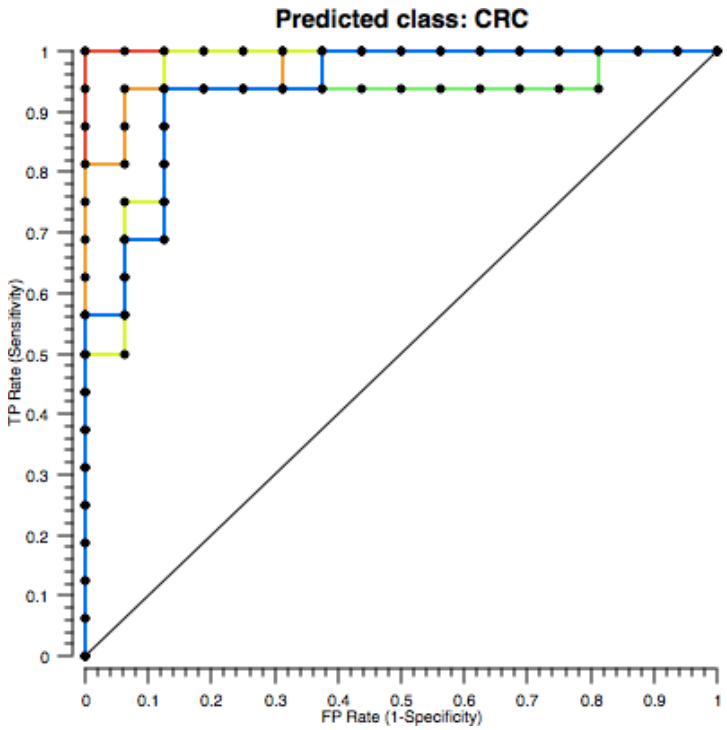


Figure 35 Benchmark-2 ROC curve for healthy-cancer using five classifiers. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The straight line represents the null model. The x-axis is the false positive rate, and the y-axis is the true positive rate. Naïve Bayes classifier performed the best on this dataset (AUC=1).

4.3 MBO1 polyp dataset

Polyp dataset sequences from the polyp-N and polyp-Y groups were preprocessed and OTU selection methods of UPARSE, UPGMA, and UCLUST were performed. Table 14 shows a summary of the preprocessing time, the resulting reads, and the number of identified OTUs for polyp dataset. The number of OTUs are 2631, 16971 and 100467 for UPARSE, UPGMA, and UCLUST respectively. As in the prior benchmarks, UPARSE returned the lowest number of OTUs and UCLUST found the highest number of OTUs for this dataset. With respect to CPU time, UPARSE is faster than the other two methods and UCLUST is significantly slower. As time is a main issue in analyzing sequences, this is one significant advantage of UPARSE. Additionally, the number of detected centroids (OTUs) are less in the UPARSE method which decreases the chance of having spurious OTUs. It is claimed that UPARSE generates OTUs that are superior to other methods including QIIME and mothur on mock community tests where the identified OTU are more accurate predictions of biological sequences and the number of OTUs are much closer to the number of known species in the community (Edgar, 2013).

Table 14 Comparing three OTU selection methods using the polyp dataset. The UPARSE was the fastest and generated the lowest number of OTUs. UCLUST was the slowest and produced the highest number of OTUs.

	UPARSE	UPGMA	UCLUST
Preprocessing and clustering time (using 30 processors)	3h40m	60h	7days
No. of original reads (for alignment)	12,646,278	12,646,278	12,646,278
No. of reads for OTU clustering	4,377,359	4,377,359	4,830,116
No. of detected OTUs	2631	16971	100467

4.3.1 MBO1 polyp dataset rarefaction

As in the prior two benchmarks, rarefaction plots were produced separately for each of three datasets to assess the sequencing depth. In Figure 36, rarefaction plots for biopsy (BS), stool swabs (HS), and rectal swabs (SS) datasets are presented. As all the curves reached a plateau, sequencing depth is acceptable, and the reads should adequately identify the number of species that are in the samples.

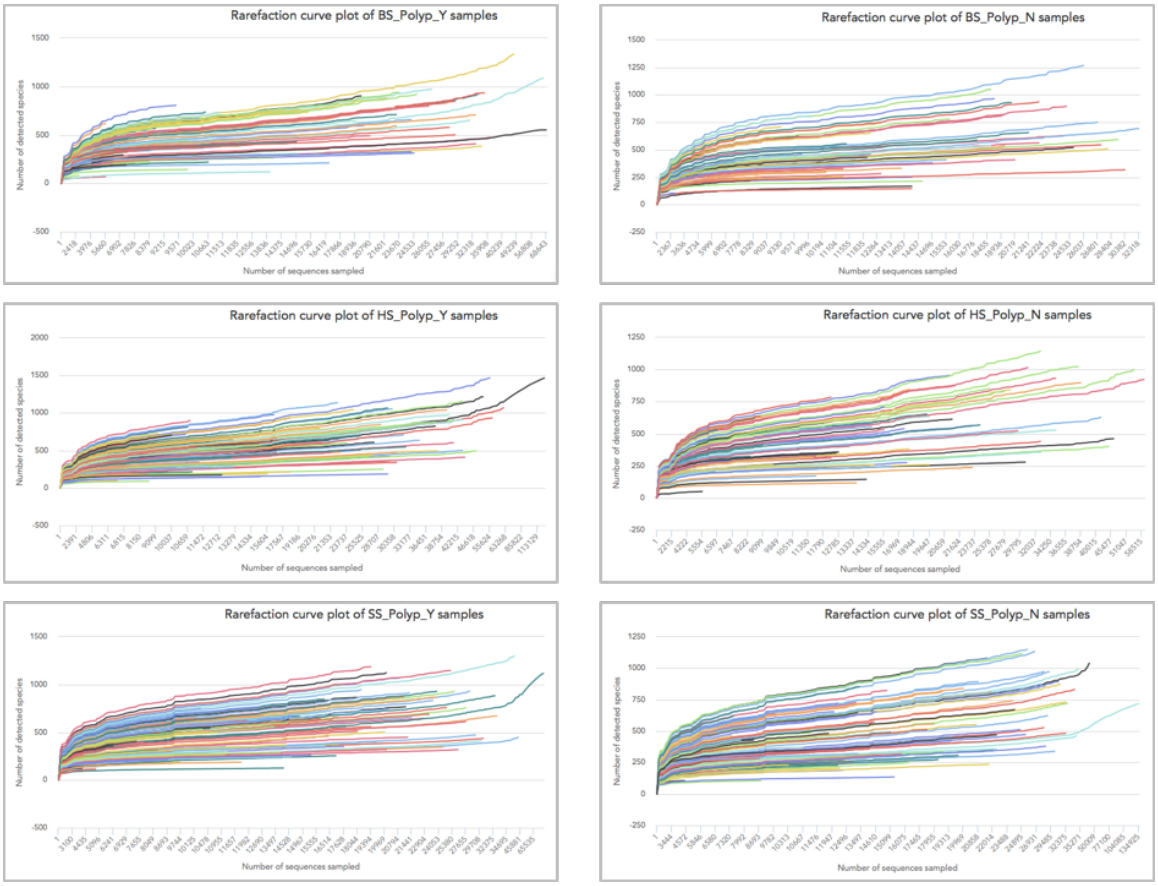


Figure 36 Polyp dataset rarefaction plots for biopsy (BS), stool (SS), and rectal swabs (HS) samples. Sequencing depth is suitable for all as the plots have reached a plateau.

4.3.2 MBO1 polyp dataset alpha diversity

The next step in the pipeline was to analyze the diversity within samples. Several diversity indices for richness and evenness were used to investigate diversity. Each of biopsy (BS), stool swabs (HS), and rectal swabs (SS) datasets were analyzed separately.

Biopsy samples alpha diversity: Alpha diversity of biopsy dataset for the polyp-Y and polyp-N groups are summarized in Table 15. The pattern of diversity is the same for the three methods, despite different absolute values. We observed higher Shannon and invSimpson as well as lower Simpson's diversity in the polyp-N group compared to the polyp-Y which indicates higher diversity in the polyp-N group. In contrast, the Sobs index shows a higher number of species in the polyp-N group using the UPARSE method. However, it is lower in the UPGMA and UCLUST method. Overall, the polyp-N group showed a higher diversity compared to the polyp-Y which may be due to shifting the bacterial composition of the colon from a healthy diverse population to a less diverse community as it responds to the disease state.

Table 15 The alpha diversity analysis of the polyp biopsy dataset using three OTU selection methods. The pattern of diversity is the same for the three methods, despite different absolute values. Overall, the polyp-N group has a higher diversity compared to the polyp-Y.

	UPARSE		UPGMA		UCLUST	
Diversity index	**Average in group**		**Average in group**		**Average in group**	
Group	**Polyp_Y**	**Polyp_N**	**Polyp_Y**	**Polyp_N**	**Polyp_Y**	**Polyp_N**
nseq	17715	16799	20994	18063	21979	19294
Shannon	3.4	3.6	4.84	4.96	4.85	5.06
Simpson	0.11	0.08	0.03	0.03	0.06	0.04
invSimpson	13.7	18.14	46	56.3	29.47	39.6
Sobs	242	251	1065	1025	1808	1729

Stool samples alpha diversity: Alpha diversity analysis for the home stool swabs is shown in Table 16. There is a very little change in the diversity of the polyp-Y and polyp-N groups based on the Shannon, Simpson and invSimpson indices. The two groups show a similar diversity pattern as we only see lower diversity in the polyp-N group compared to the polyp-Y for the sobs index.

Table 16 The alpha diversity analysis of the polyp stool dataset using three OTU selection methods. There is a slight change in the diversity of polyp-Y and polyp-N groups.

	UPARSE		UPGMA		UCLUST	
Diversity index	**Average in group**		**Average in group**		**Average in group**	
Group	**Polyp_Y**	**Polyp_N**	**Polyp_Y**	**Polyp_N**	**Polyp_Y**	**Polyp_N**
nseq	22136	20495	25922	22394	27259	23493
Shannon	3.5	3.48	4.84	4.82	4.95	4.94
Simpson	0.08	0.09	0.03	0.03	0.04	0.04
invSimpson	16.9	16.3	50	49.9	39.38	39.17
Sobs	242	230	1112	983	1944	1715

Rectal Swab alpha diversity: The alpha diversity results for the rectal swabs dataset is shown in Table 17. The diversity pattern is the same for all three OTU selection methods. The diversity in the polyp-N and polyp-Y groups is similar to the stool dataset where no change in diversity was seen based on the Shannon, Simpson, and invSimpson metrics. There is a lower sobs index in the polyp-N group compared to the polyp-Y group.

Table 17 The alpha diversity analysis of the rectal swabs using three OTU selection methods. There is a slight change in the diversity of the polyp-Y and polyp-N groups.

	UPARSE		UPGMA		UCLUST	
Diversity index	**Average in group**		**Average in group**		**Average in group**	
Group	**Polyp-Y**	**Polyp-N**	**Polyp-Y**	**Polyp-N**	**Polyp-Y**	**Polyp-N**
nseq	22136	20495	25922	22394	27259	23493
Shannon	3.5	3.48	4.84	4.82	4.95	4.94
Simpson	0.08	0.09	0.03	0.03	0.04	0.04
invSimpson	16.9	16.3	50	49.9	39.38	39.17
Sobs	242	230	1112	983	1944	1715

4.3.3 MBO1 polyp dataset beta diversity

As described for the prior Benchmarks, the phylogenetic UniFrac metric and nonphylogenetic metric Bray-Curtis was used for beta diversity analysis. It is evident from UniFrac PCoA plots in Figure 37, Figure 38, and Figure 39 in addition to Bray-Curtis PCoA in Figure 40 that neither the UniFrac nor the Bray-Curtis metrics successfully separated the polyp-Y and polyp-N groups from each other.

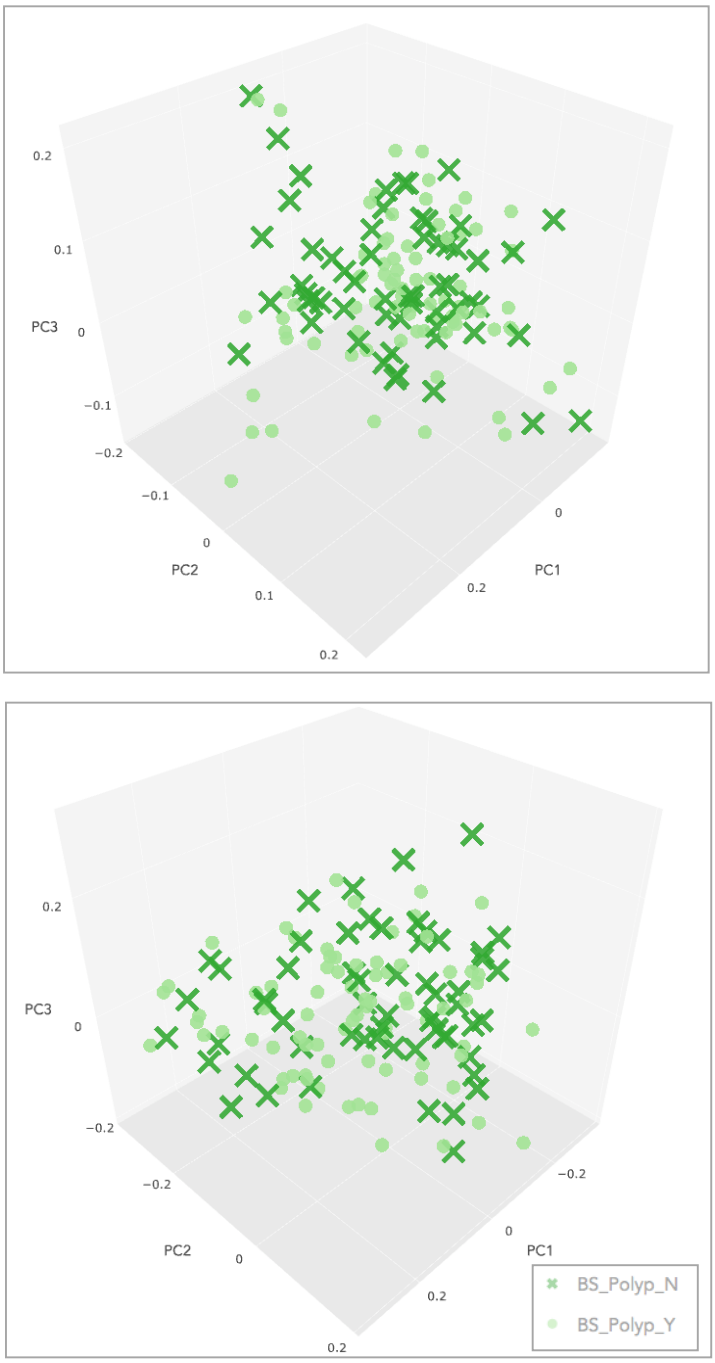


Figure 37 UniFrac PCoA plots of the biopsy (BS) dataset.
Neither unweighted (top) nor weighted (bottom) UniFrac were successful in discriminating polyp-Y and polyp-N groups. The axes are principal components.

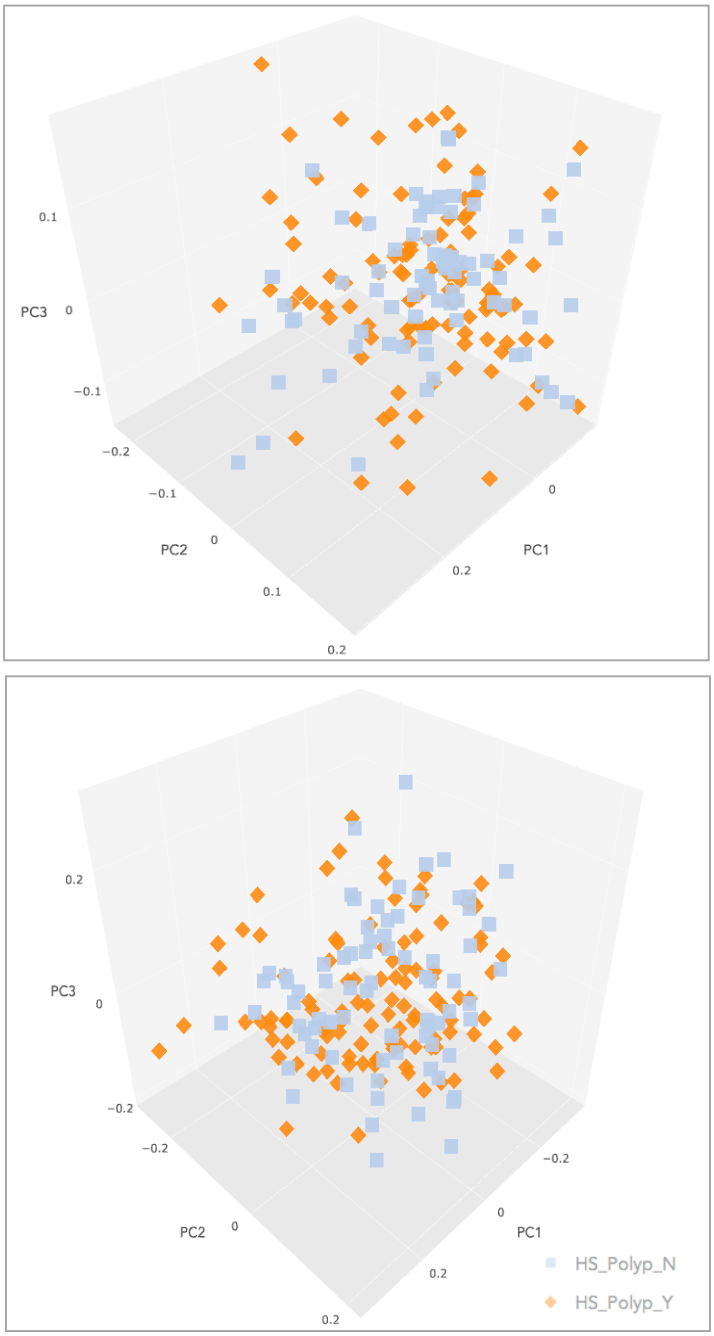


Figure 38 UniFrac PCoA plots of the stool swabs (HS) dataset.
Neither unweighted (top) nor weighted (bottom) UniFrac were successful in discriminating polyp-Y and polyp-N groups. The axes are principal components.

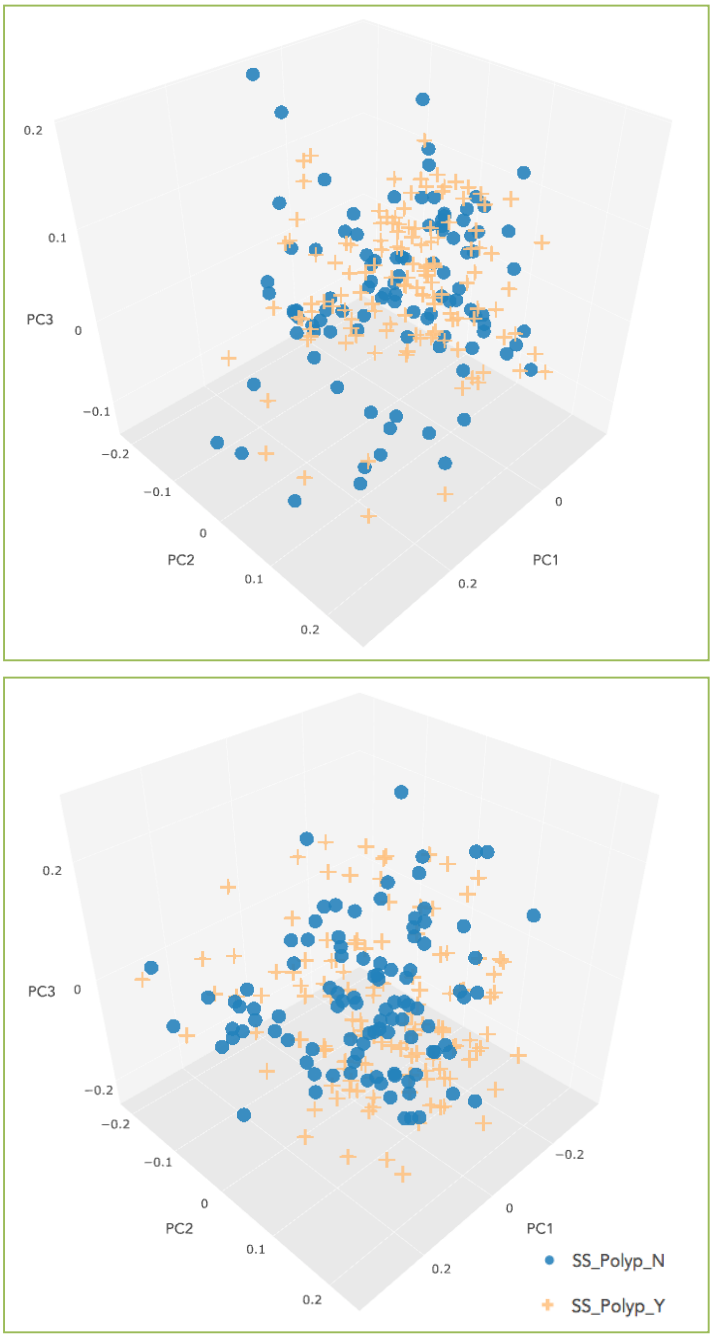


Figure 39 UniFrac PCoA plots of the rectal swabs (SS) dataset.
Neither unweighted (top) nor weighted (bottom) UniFrac were successful in discriminating polyp-Y and polyp-N groups. The axes are principal components.

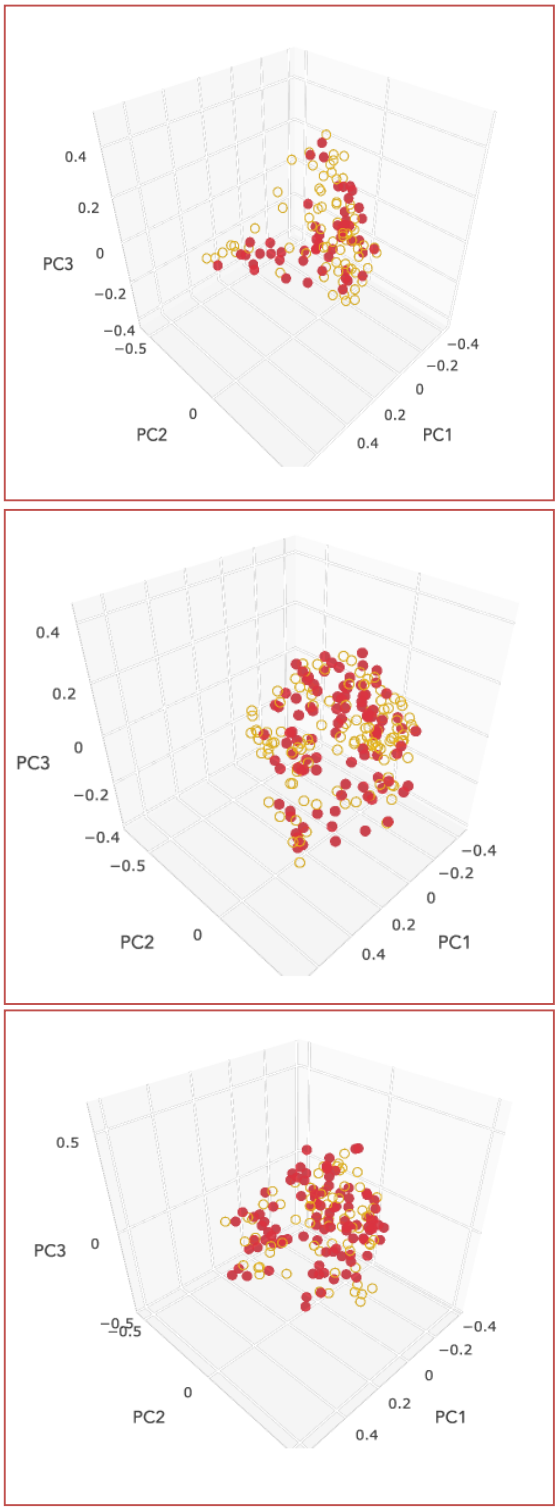


Figure 40 Bray-Curtis PCoA plots of the biopsy (BS), stool samples (HS), and rectal swabs (SS) from the polyp dataset.
No dissimilarity detected between the polyp-Y and polyp-N groups using this metric. The axes are principal components.

4.3.4 MBO1 polyp dataset significant OTUs

In a comparison of the polyp-Y and polyp-N groups, significantly different OTUs ($P < 0.05$) were detected with Kruskal-Wallis, MetaStats, LEfSe, and Indicator analysis. We retained all significant OTUs and used them for classification as in Benchmark-1 and Benchmark-2. Table 18 shows the number of retained significant OTUs for the polyp dataset. Similar to the number of OTUs, the number of significant OTUs was also the lowest with the UPARSE method, and the significant OTUs was highest with UCLUST.

Table 18 Number of significant OTUs detected in each clustering method for any of the specimen types in the polyp dataset.
The number of significant OTUs are lower when the number of total OTUs are lower.

	Specimen	UPARSE	UPGMA	UCLUST
No. of total OTUs	all	2631	16971	100467
No. of significant OTUs	Biopsy	109	291	449
No. of significant OTUs	Stool	59	254	358
No. of significant OTUs	Swab	92	396	584

Significant OTUs with the biopsy, rectal swabs, and stool samples are listed in Table 19, Table 20, and Table 21, respectively. There are significant OTUs at each taxonomic level and, as these OTUs are significantly different between the polyp-Y and polyp-N, we can use them as classification features to improve the classifiers' performance. Bar plots of the normalized abundance of the significant OTUs are presented in Figure 41, Figure 42, and Figure 43 for biopsy, swab, and stool samples, respectively. As in the

previous benchmarks, some taxa are enriched while others are depleted in the polyp-Y group. However, some OTUs have the same pattern of change in all three sample types. For example, *Bacteroides* is enriched in the polyp-Y group in the biopsy, stool, and swab samples. In contrast, there are some OTUs with a different pattern in different sample types. For instance, *Blautia* shows enrichment in the polyp-Y stool and swab samples and depletion in the polyp-Y biopsies. As biopsy samples are representing mucosal microbes, it is not surprising that the OTUs and the direction of their disease associations differ with biopsies. There were some OTUs that show different trends between swab and stool samples, such as *Faecalibacterium* which is increased in the polyp-Y swabs and decreased in the polyp-Y stool samples. Sampling confounders could be responsible for this difference. For example, the date of sample collection may play a role as biopsies, stool samples, and rectal swabs were not necessarily collected on the same day, e.g., the microbiome could change between samplings due to recent dietary changes. Another parameter that could be responsible is the collection site of the samples, as stool is the representative of the luminal bacteria from the entire colon while swab samples may carry both luminal and mucosal bacteria from the last segment of colon. There is also a chance that there are technical processing errors that may cause the problem. For example, difficulties in sample preparation, sequencing, and sequence analysis may have confounded the results.

Table 19 The significantly different OTUs between the polyp-Y and polyp-N groups in the biopsy samples. From 109 OTUs, 83 of them belong to Firmicutes phylum, and 12 OTUs were classified as Bacteroidetes Phylum. Ten OTUs have taxonomies up to species level. P: phylum; c: class; o: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU1005	GG349067	p_Firmicutes;c_Clostridia;o_Clostridiales;
2	OTU102	GG1067865	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g_Burkholderia;
3	OTU1062	GG4315468	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
4	OTU1063	GG1522739	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
5	OTU1066	GG214036	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;
6	OTU1075	GG739971	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
7	OTU1130	GG213870	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
8	OTU115	GG359930	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;s_obeum
9	OTU1177	GG517201	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
10	OTU1202	GG4367656	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Moryella;
11	OTU1206	GG183845	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
12	OTU1226	GG174752	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
13	OTU1237	GG3940440	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
14	OTU124	GG198054	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Eubacterium;s_dolichum
15	OTU13	GG366744	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
16	OTU133	GG181572	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
17	OTU1330	GG189667	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea;
18	OTU1363	GG338730	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
19	OTU1387	GG265234	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
20	OTU140	GG3199564	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Odoribacteraceae;g_Odoribacter;
21	OTU1421	GG4316515	p_Firmicutes;c_Clostridia;o_Clostridiales;
22	OTU1439	GG194172	p_Firmicutes;c_Clostridia;o_Clostridiales;
23	OTU147	GG584978	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
24	OTU1490	GG608244	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
25	OTU1519	GG589032	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
26	OTU1520	GG187741	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
27	OTU1533	GG520827	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_S24-7;
28	OTU154	GG332027	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
29	OTU1546	GG296589	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
30	OTU155	GG313166	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
31	OTU1550	GG4436120	p_Firmicutes;c_Clostridia;o_Clostridiales;
32	OTU156	GG4144206	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;

No.	OTU Number	GreenGenes ID	Taxonomy
33	OTU1595	GG225338	p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Aggregatibacter;
34	OTU1612	GG297160	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
35	OTU1644	GG216862	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
36	OTU1649	GG3804871	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
37	OTU1662	GG210793	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
38	OTU1681	GG849346	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
39	OTU1700	GG1098709	p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_Staphylococcus;
40	OTU173	GG4321260	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Paraprevotellaceae;g_Prevotella;
41	OTU1743	GG2924870	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
42	OTU1747	GG725212	p_Firmicutes;c_Clostridia;o_Clostridiales;
43	OTU177	GG195088	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
44	OTU1802	GG573969	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
45	OTU1813	GG210303	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Odoribacteraceae;g_Odoribacter;
46	OTU1886	GG320055	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
47	OTU1950	GG331253	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
48	OTU1975	GG519882	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
49	OTU198	GG988932	p_Firmicutes;c_Clostridia;o_Clostridiales;
50	OTU2020	GG229348	p_Firmicutes;c_Clostridia;o_Clostridiales;
51	OTU2058	GG4311934	p_Firmicutes;c_Clostridia;o_Clostridiales;
52	OTU2185	GG560873	p_Firmicutes;c_Clostridia;o_Clostridiales;
53	OTU2194	GG1064335	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;
54	OTU241	GG256015	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
55	OTU252	GG198118	p_Firmicutes;c_Clostridia;o_Clostridiales;
56	OTU2522	GG194097	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
57	OTU257	GG366987	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
58	OTU271	GG4387453	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
59	OTU284	GG184217	p_Firmicutes;c_Clostridia;o_Clostridiales;
60	OTU318	GG2617854	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
61	OTU320	GG1097287	p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_Staphylococcus;s_epidermidis
62	OTU334	GG181826	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
63	OTU366	GG516265	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;
64	OTU384	GG16274	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_Clostridium;
65	OTU394	GG198085	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
66	OTU398	GG192155	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
67	OTU404	GG308631	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;

No.	OTU Number	GreenGenes ID	Taxonomy
68	OTU415	GG324015	p_Firmicutes;c_Clostridia;o_Clostridiales;
69	OTU436	GG1024958	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Rothia;s_aeria
70	OTU444	GG204236	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
71	OTU449	GG198937	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Holdemania;
72	OTU461	GG359359	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
73	OTU474	GG330439	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;s_stercorea
74	OTU475	GG354574	p_Proteobacteria;c_Deltaproteobacteria;o_Desulfovibrionales;f_Desulfovibrionaceae;g_Bilophila;
75	OTU51	GG1003657	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Klebsiella;
76	OTU514	GG848116	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
77	OTU550	GG2134456	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
78	OTU553	GG1135793	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Peptococcaceae;g_Peptococcus;
79	OTU555	GG199350	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
80	OTU567	GG585220	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_cc_115;
81	OTU571	GG187802	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;
82	OTU576	GG218102	p_Firmicutes;c_Clostridia;o_Clostridiales;
83	OTU6	GG1111141	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
84	OTU600	GG2042960	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Peptoniphilus;
85	OTU612	GG531950	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
86	OTU647	GG1007247	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_Granulicatella;
87	OTU649	GG353631	p_Firmicutes;c_Clostridia;o_Clostridiales;
88	OTU65	GG305288	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_Clostridium;
89	OTU655	GG4315787	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
90	OTU659	GG176104	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
91	OTU662	GG179815	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;s_gnavus
92	OTU697	GG4303851	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
93	OTU723	GG230405	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
94	OTU737	GG4470076	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;s_nigrescens
95	OTU76	GG851704	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Parvimonas;
96	OTU761	GG358030	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
97	OTU791	GG4453550	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_RFN20;
98	OTU812	GG191100	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
99	OTU82	GG192424	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
100	OTU83	GG882886	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_fragilis
101	OTU838	GG20310	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;

No.	OTU Number	GreenGenes ID	Taxonomy
102	OTU844	GG4440820	p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Leptotrichiaceae;g_Leptotrichia;
103	OTU85	GG843886	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
104	OTU861	GG4329995	p_Proteobacteria;c_Deltaproteobacteria;o_Desulfovibrionales;f_Desulfovibrionaceae;
105	OTU874	GG932843	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Anaerococcus;
106	OTU899	GG105514	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
107	OTU905	GG575852	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Citrobacter;
108	OTU914	GG1835985	p_Firmicutes;c_Clostridia;o_Clostridiales;
109	OTU969	GG574988	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;

Table 20 The significantly different OTUs between the polyp-Y and polyp-N groups in the stool samples. From 59 significant OTUs, 42 of them belong to Firmicutes phylum, 7 OTUs to Bacteroidetes, phylum. The taxonomy of Five OTUs is clear up to species level. P: phylum; c: class; o: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU1002	GG915327	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_Granulicatella;
2	OTU1159	GG295456	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;s_producta
3	OTU1166	GG727205	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_Lactobacillus;s_zeae
4	OTU12	GG260579	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;g_Alistipes;s_putredinis
5	OTU125	GG3443119	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
6	OTU126	GG3902153	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Barnesiellaceae;
7	OTU1353	GG1904686	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Eubacterium;s_dolichum
8	OTU1374	GG4332878	p_Actinobacteria;c_Coriobacteriia;o_Coriobacteriales;f_Coriobacteriaceae;g_Slackia;
9	OTU1387	GG265234	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
10	OTU147	GG584978	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
11	OTU1476	GG851812	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Eubacteriaceae;g_Anaerofustis;
12	OTU1520	GG187741	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
13	OTU1576	GG121873	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Dehalobacteriaceae;g_Dehalobacterium;
14	OTU1608	GG234447	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
15	OTU1774	GG260691	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Allobaculum;
16	OTU1899	GG348215	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
17	OTU1942	GG197249	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
18	OTU2022	GG2782816	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;
19	OTU2159	GG4366089	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
20	OTU2221	GG240240	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Paraprevotellaceae;g_Prevotella;
21	OTU2427	GG198720	p_Firmicutes;c_Clostridia;o_Clostridiales;
22	OTU243	GG362380	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
23	OTU2495	GG198122	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
24	OTU2525	GG2990918	p_Actinobacteria;c_Coriobacteriia;o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella;s_stercoris
25	OTU258	GG198569	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
26	OTU2637	GG1079013	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Corynebacteriaceae;g_Corynebacterium;
27	OTU274	GG1083336	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Rothia;s_mucilaginosa
28	OTU305	GG197517	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Barnesiellaceae;
29	OTU319	GG4397094	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;
30	OTU324	GG4327303	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Lachnospira;
31	OTU327	GG3708846	p_Firmicutes;c_Clostridia;o_Clostridiales;
32	OTU346	GG1658654	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Epulopiscium;
33	OTU35	GG851668	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;

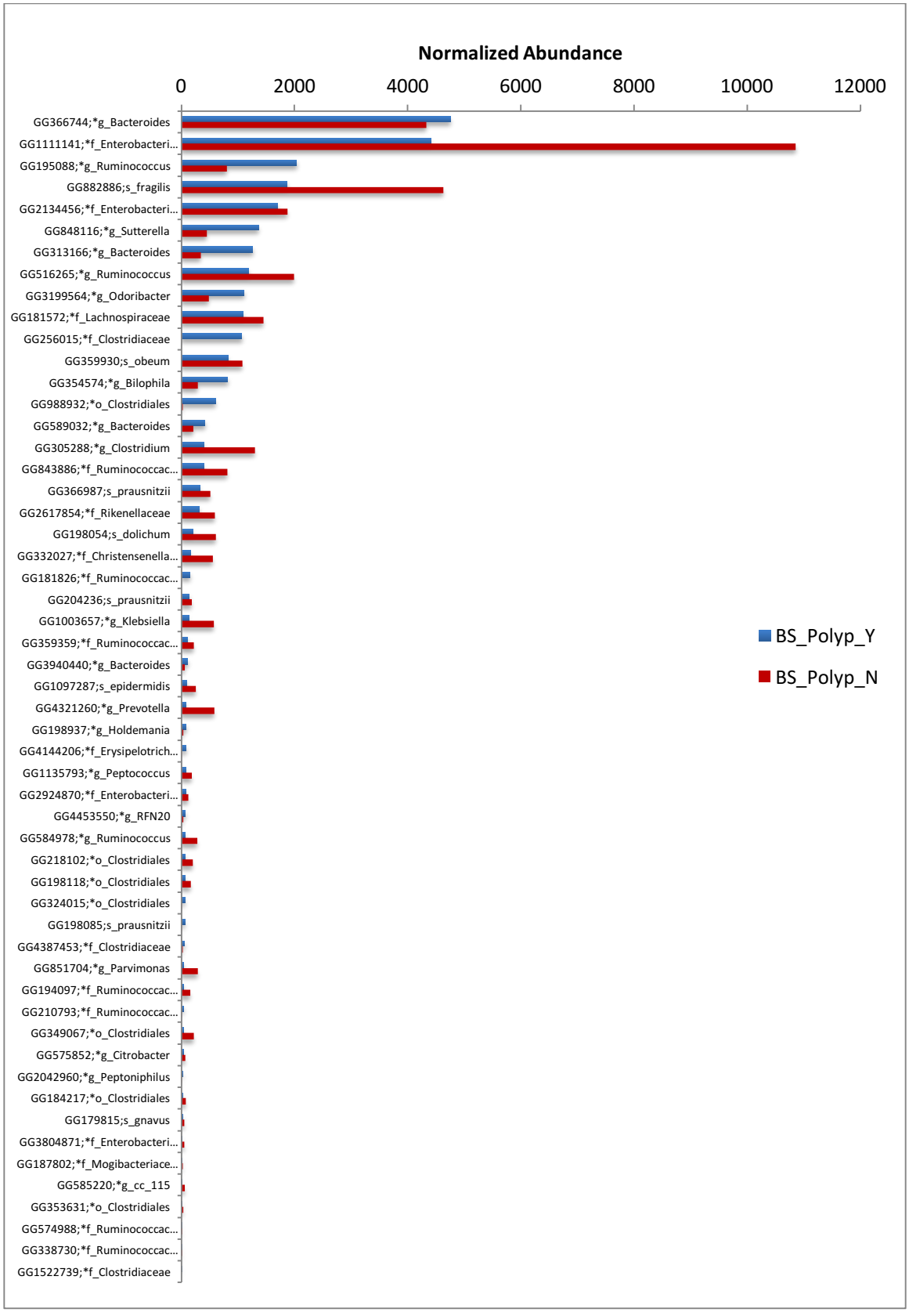
No.	OTU Number	GreenGenes ID	Taxonomy
34	OTU352	GG246930	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister;
35	OTU389	GG208088	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Odoribacteraceae;g_Butyricimonas;
36	OTU409	GG1667433	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea;
37	OTU434	GG966508	p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Haemophilus;
38	OTU440	GG177310	p_Firmicutes;c_Clostridia;o_Clostridiales;
39	OTU443	GG198453	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
40	OTU453	GG2986828	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
41	OTU517	GG535901	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister;
42	OTU535	GG195807	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
43	OTU560	GG804624	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Peptostreptococcaceae;g_Peptostreptococcus;
44	OTU576	GG218102	p_Firmicutes;c_Clostridia;o_Clostridiales;
45	OTU593	GG434992	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Peptostreptococcaceae;
46	OTU636	GG4336939	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
47	OTU647	GG1007247	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_Granulicatella;
48	OTU675	GG893214	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Mobiluncus;
49	OTU733	GG1050844	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Micrococcus;s_luteus
50	OTU747	GG560336	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
51	OTU785	GG192465	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
52	OTU786	GG951711	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;
53	OTU790	GG560873	p_Firmicutes;c_Clostridia;o_Clostridiales;
54	OTU808	GG4314545	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinobaculum;
55	OTU819	GG4364747	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
56	OTU847	GG370809	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
57	OTU96	GG609964	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Peptoniphilus;
58	OTU960	GG298995	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
59	OTU964	GG352914	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii

Table 21 The significantly different OTUs between the polyp-Y and polyp-N groups in the swab samples. From 92 significant OTUs, 69 OTUs belong to Firmicutes, and 12 OTUs belong to Bacteroidetes Phyla. The taxonomy of 10 OTUs is clear up to species level. P: phylum; c: class; o: order; f: family; g: genus; s: species.

No.	OTU Number	GreenGenes ID	Taxonomy
1	OTU1001	GG4456576	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
2	OTU1014	GG178478	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
3	OTU1033	GG189606	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
4	OTU106	GG366515	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister;
5	OTU1071	GG174885	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
6	OTU1098	GG4388068	p_Tenericutes;c_Mollicutes;o_RF39;
7	OTU112	GG187668	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_Clostridium;
8	OTU116	GG314996	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
9	OTU1164	GG994357	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;g_Mogibacterium;
10	OTU1167	GG359788	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
11	OTU117	GG963388	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
12	OTU122	GG248563	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;
13	OTU1298	GG213870	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
14	OTU1453	GG146086	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
15	OTU149	GG1057169	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella;
16	OTU1513	GG3940412	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
17	OTU154	GG332027	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Christensenellaceae;
18	OTU1558	GG269455	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
19	OTU158	GG776472	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Enterococcaceae;g_Enterococcus;
20	OTU1584	GG185066	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
21	OTU1611	GG3326658	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
22	OTU1633	GG948369	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Streptococcus;
23	OTU1700	GG1098709	p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_Staphylococcus;
24	OTU1707	GG208377	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
25	OTU1732	GG4387706	p_Firmicutes;c_Clostridia;o_Clostridiales;
26	OTU175	GG4409213	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
27	OTU1781	GG358342	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
28	OTU193	GG367113	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
29	OTU1942	GG197249	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;
30	OTU198	GG988932	p_Firmicutes;c_Clostridia;o_Clostridiales;
31	OTU1992	GG190299	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Lachnospira;
32	OTU2	GG588308	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
33	OTU2001	GG526682	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces;

No.	OTU Number	GreenGenes ID	Taxonomy
34	OTU2068	GG527988	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Eubacteriaceae;g_Anaerofustis;
35	OTU226	GG358265	p_Firmicutes;c_Clostridia;o_Clostridiales;
36	OTU2283	GG348006	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;s_lactaris
37	OTU2378	GG364516	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
38	OTU2413	GG3898650	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
39	OTU2447	GG312677	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
40	OTU2536	GG1072223	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Streptococcus;s_infantis
41	OTU2565	GG589032	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
42	OTU27	GG4389944	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_ph2;
43	OTU272	GG363830	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
44	OTU286	GG199668	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
45	OTU304	GG369970	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;
46	OTU31	GG213810	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Roseburia;
47	OTU318	GG2617854	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
48	OTU325	GG1028501	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Dialister;
49	OTU328	GG50025	p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium;s_adolescentis
50	OTU329	GG214031	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;
51	OTU331	GG368448	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
52	OTU336	GG62513	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Moryella;s_indoligenes
53	OTU362	GG966186	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Varibaculum;
54	OTU366	GG516265	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;
55	OTU368	GG204093	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
56	OTU398	GG192155	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
57	OTU436	GG1024958	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Rothia;s_aeria
58	OTU444	GG204236	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium;s_prausnitzii
59	OTU449	GG198937	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Holdemania;
60	OTU45	GG581003	p_Firmicutes;c_Clostridia;o_Clostridiales;
61	OTU453	GG2986828	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
62	OTU454	GG2202001	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
63	OTU457	GG358613	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
64	OTU469	GG383714	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Anaerococcus;
65	OTU470	GG1052809	p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Varibaculum;
66	OTU473	GG192210	p_Firmicutes;c_Clostridia;o_Clostridiales;
67	OTU496	GG342947	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;

No.	OTU Number	GreenGenes ID	Taxonomy
68	OTU5	GG368935	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
69	OTU50	GG529744	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Bulleidia;s_moorei
70	OTU51	GG1003657	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Klebsiella;
71	OTU520	GG659221	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;
72	OTU53	GG367044	p_Firmicutes;c_Clostridia;o_Clostridiales;
73	OTU550	GG2134456	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
74	OTU555	GG199350	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
75	OTU59	GG577294	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides;s_distasonis
76	OTU6	GG1111141	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;
77	OTU628	GG2047910	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides;
78	OTU63	GG363343	p_Firmicutes;c_Clostridia;o_Clostridiales;
79	OTU632	GG217047	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Tissierellaceae;g_Anaerococcus;
80	OTU672	GG362765	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
81	OTU686	GG327149	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
82	OTU720	GG175910	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
83	OTU739	GG303794	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
84	OTU768	GG1820776	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
85	OTU77	GG1985	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_ovatus
86	OTU772	GG363467	p_Firmicutes;c_Clostridia;o_Clostridiales;
87	OTU846	GG186654	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
88	OTU849	GG190226	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
89	OTU872	GG561595	p_Actinobacteria;c_Coriobacteriia;o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella;s_aerofaciens
90	OTU889	GG320321	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
91	OTU929	GG216010	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Mogibacteriaceae;
92	OTU966	GG536281	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus;


Normalized Abundance
0
2000
4000
6000
8000
10000
12000
GG366744;*g_Bacteroides
GG1111141;*f_Enterobacteri...
GG195088;*g_Ruminococcus
GG882886;s_fragilis
GG2134456;*f_Enterobacteri...
GG848116;*g_Sutterella
GG313166;*g_Bacteroides
GG516265;*g_Ruminococcus
GG3199564;*g_Odoribacter
GG181572;*f_Lachnospiraceae
GG256015;*f_Clostridiaceae
GG359930;s_obeum
GG354574;*g_Bilophila
GG988932;*o_Clostridiales
GG589032;*g_Bacteroides
GG305288;*g_Clostridium
GG843886;*f_Ruminococcac...
GG366987;s_prausnitzii
GG2617854;*f_Rikenellaceae
GG198054;s_dolichum
GG332027;*f_Christensenella...
GG181826;*f_Ruminococcac...
GG204236;s_prausnitzii
GG1003657;*g_Klebsiella
GG359359;*f_Ruminococcac...
GG3940440;*g_Bacteroides
GG1097287;s_epidermidis
GG4321260;*g_Prevotella
GG198937;*g_Holdemania
GG4144206;*f_Erysipelotrich...
GG1135793;*g_Peptococcus
GG2924870;*f_Enterobacteri...
GG4453550;*g_RFN20
GG584978;*g_Ruminococcus
GG218102;*o_Clostridiales
GG198118;*o_Clostridiales
GG324015;*o_Clostridiales
GG198085;s_prausnitzii
GG4387453;*f_Clostridiaceae
GG851704;*g_Parvimonas
GG194097;*f_Ruminococcac...
GG210793;*f_Ruminococcac...
GG349067;*o_Clostridiales
GG575852;*g_Citrobacter
GG2042960;*g_Peptoniphilus
GG184217;*o_Clostridiales
GG179815;s_gnavus
GG3804871;*f_Enterobacteri...
GG187802;*f_Mogibacteriace...
GG585220;*g_cc_115
GG353631;*o_Clostridiales
GG574988;*f_Ruminococcac...
GG338730;*f_Ruminococcac...
GG1522739;*f_Clostridiaceae
BS_Polyp_Y
BS_Polyp_N

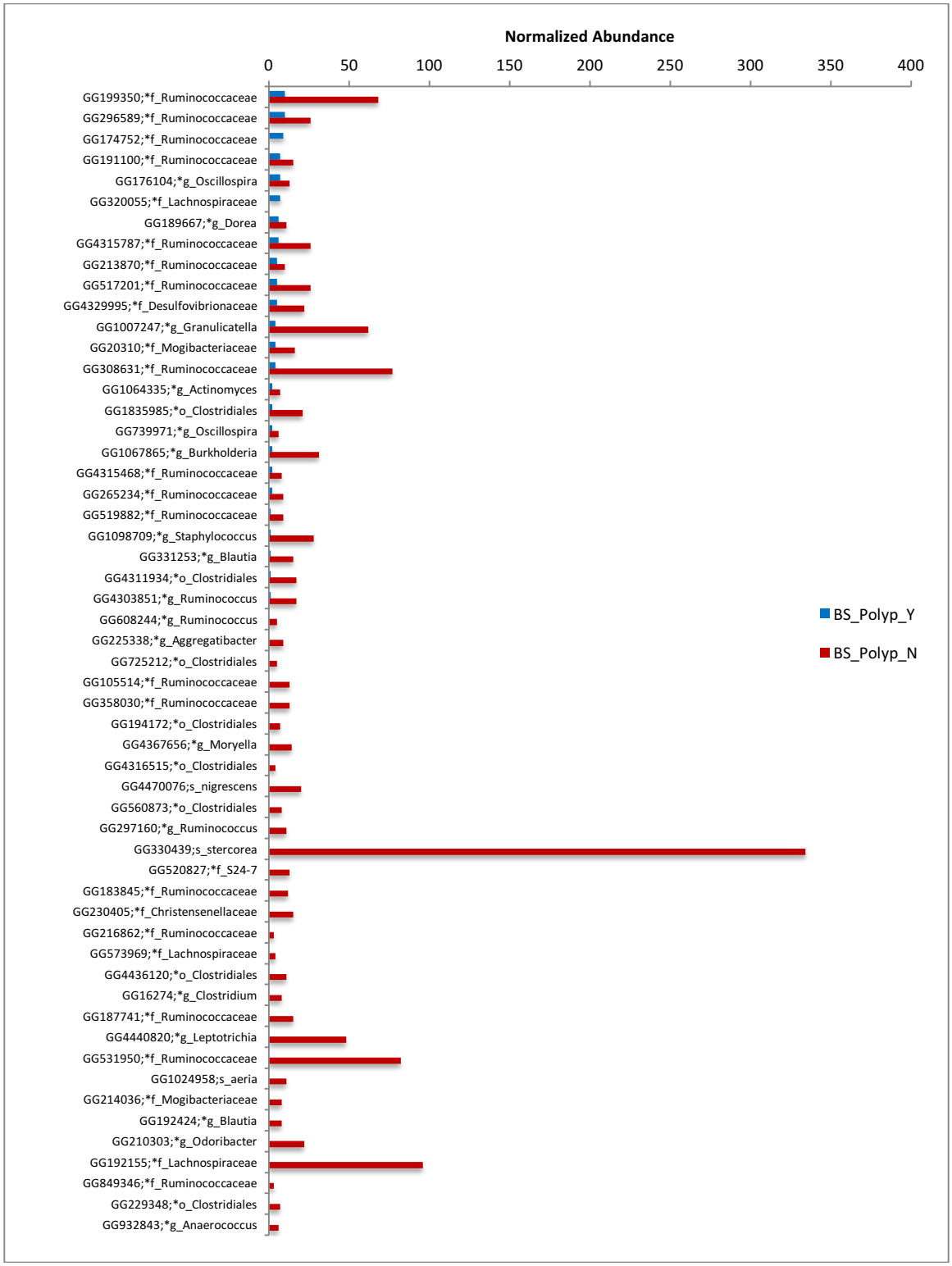


Figure 41 Change of bacterial taxa in the polyp-Y and polyp-N in biopsy samples (BS). Some taxa increased, and some decreased in the polyp-Y group compared to the polyp-N.

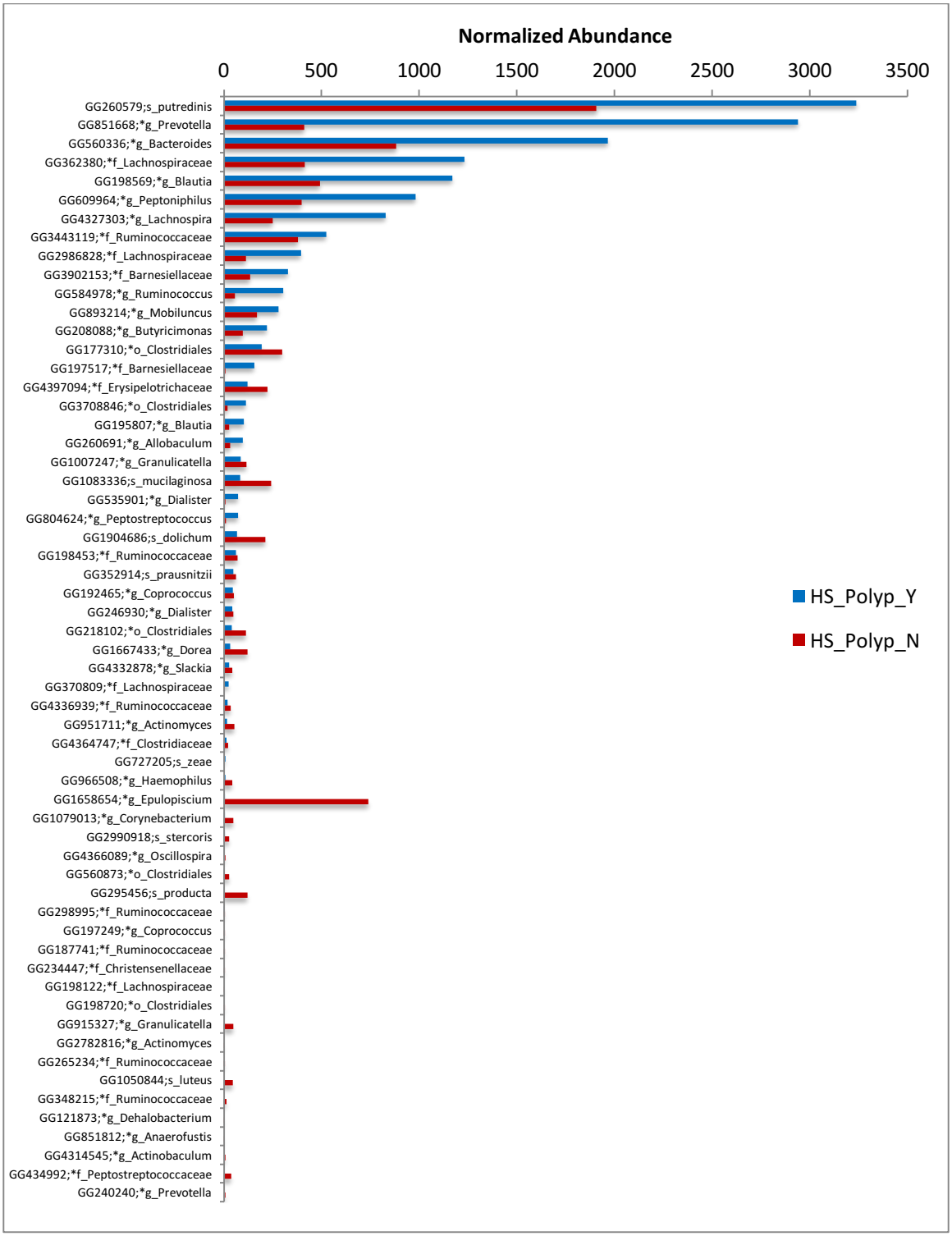
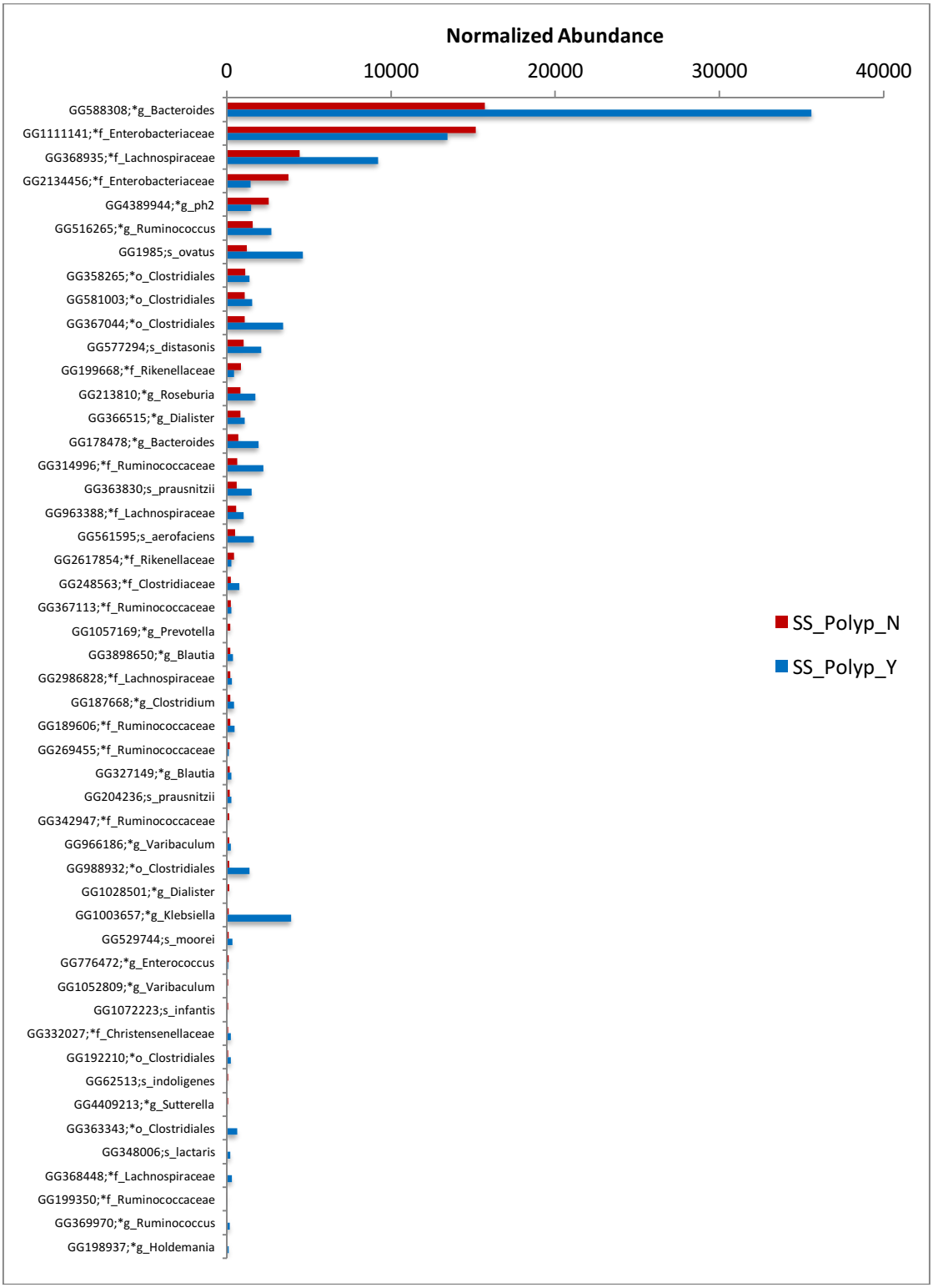


Figure 42 Change of bacterial taxa in the polyp-Y and polyp-N in home stool samples (HS). Some taxa increased, and some decreased in the polyp-Y group compared to the polyp-N.


Normalized Abundance
0
10000
20000
30000
40000
GG588308;*g_Bacteroides
GG1111141;*f_Enterobacteriaceae
GG368935;*f_Lachnospiraceae
GG2134456;*f_Enterobacteriaceae
GG4389944;*g_ph2
GG516265;*g_Ruminococcus
GG1985;s_ovatus
GG358265;*o_Clostridiales
GG581003;*o_Clostridiales
GG367044;*o_Clostridiales
GG577294;s_distasonis
GG199668;*f_Rikenellaceae
GG213810;*g_Roseburia
GG366515;*g_Dialister
GG178478;*g_Bacteroides
GG314996;*f_Ruminococcaceae
GG363830;s_prausnitzii
GG963388;*f_Lachnospiraceae
GG561595;s_aerofaciens
GG2617854;*f_Rikenellaceae
GG248563;*f_Clostridiaceae
GG367113;*f_Ruminococcaceae
GG1057169;*g_Prevotella
GG3898650;*g_Blautia
GG2986828;*f_Lachnospiraceae
GG187668;*g_Clostridium
GG189606;*f_Ruminococcaceae
GG269455;*f_Ruminococcaceae
GG327149;*g_Blautia
GG204236;s_prausnitzii
GG342947;*f_Ruminococcaceae
GG966186;*g_Varibaculum
GG988932;*o_Clostridiales
GG1028501;*g_Dialister
GG1003657;*g_Klebsiella
GG529744;s_moorei
GG776472;*g_Enterococcus
GG1052809;*g_Varibaculum
GG1072223;s_infantis
GG332027;*f_Christensenellaceae
GG192210;*o_Clostridiales
GG62513;s_indoligenes
GG4409213;*g_Sutterella
GG363343;*o_Clostridiales
GG348006;s_lactaris
GG368448;*f_Lachnospiraceae
GG199350;*f_Ruminococcaceae
GG369970;*g_Ruminococcus
GG198937;*g_Holdemania
SS_Polyp_N
SS_Polyp_Y

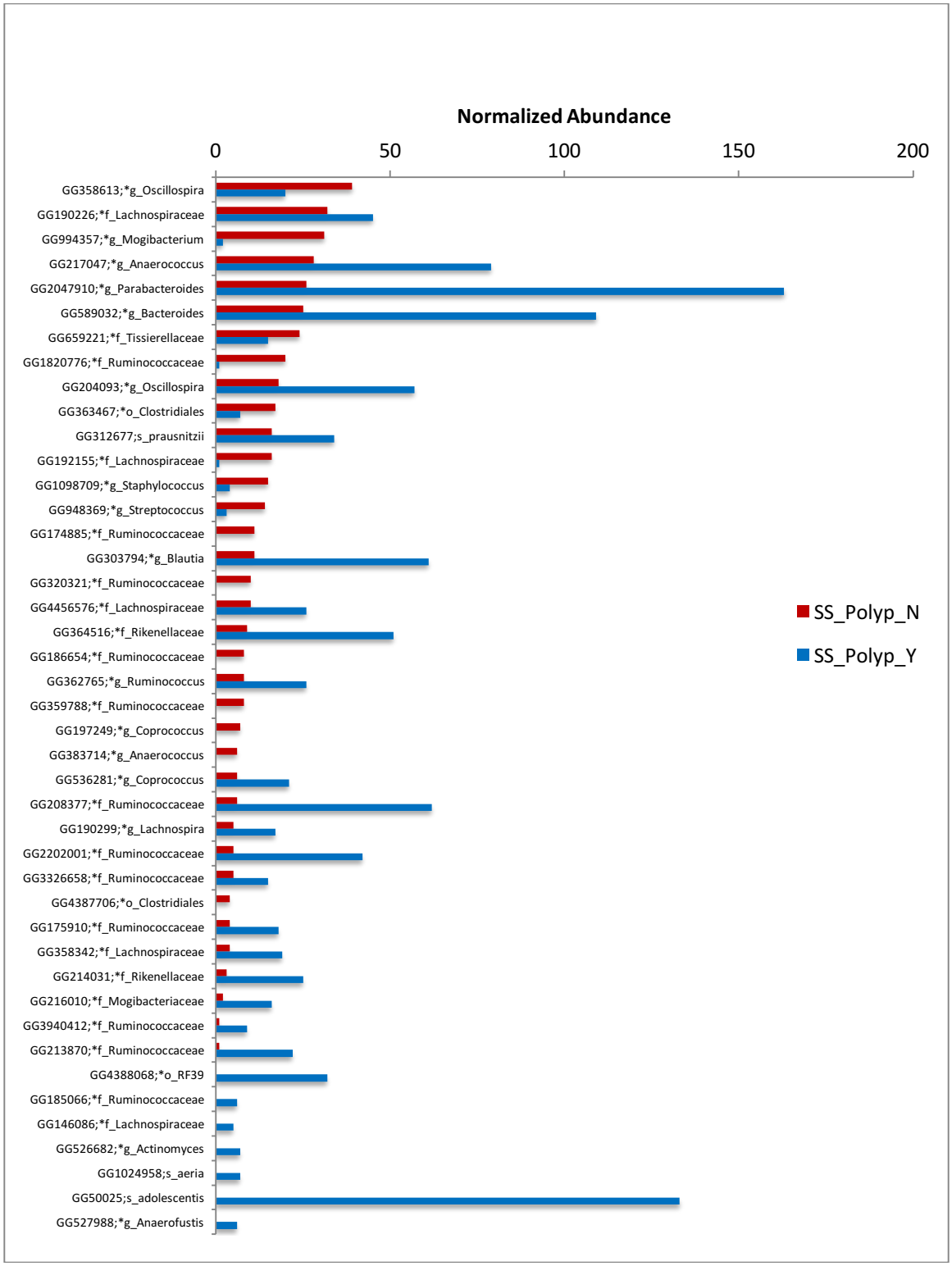


Figure 43 Change of bacterial taxa in the polyp-Y and polyp-N in rectal swabs (SS). Some taxa increased, and some decreased in the polyp-Y group compared to the polyp-N.

4.3.5 MBO1 polyp dataset classification

The relative abundance tables from each method were rarefied and used for data mining using the orange data mining pipeline. Classification methods were performed separately for the biopsy, stool, and rectal swabs datasets. Using the 5-fold cross-validation method, 80% of samples were chosen as the training set, and the rest were used as the test set. The same data mining procedure was done for both abundance tables with all features (raw) and significant features (filtered) as described above for Benchmark-1 and 2.

4.3.6 MBO1 polyp dataset classification validation

The classifiers were validated using 10-fold cross-validation method. Results of validation for the UPARSE method are shown in Table 22. In all specimen types (biopsy, rectal swab, and stool samples), the classification accuracy, sensitivity, specificity, and area under ROC curve improved by using significant OTUs for classification instead of all OTUs. This demonstrates that statistically significant features can be more informative for predicting classes (polyp-N or polyp-Y).

Table 22 The classification validation results of MBO1 polyp dataset.
The top tables show validation results of the classifiers produced using all OTUs. The middle tables used just the significant OTUs as classification features. The bottom is the results of applying the second classifier on the test dataset. Classifiers are performing better when significant OTUs are used as classification features.

Polyp_UPARSE

Raw_BS				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.43	0	1	0.59
Random Forest	0.63	0.77	0.44	0.66
kNN	0.55	0.66	0.39	0.61
Classification Tree	0.56	0.66	0.41	0.57
Logistic regression	0.56	0.59	0.51	0.57
Neural Network	0.56	0.63	0.46	0.57
SVM	0.57	1	0	0.50

Raw_HS				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.40	0	1	0.57
Random Forest	0.55	0.78	0.23	0.56
kNN	0.58	0.87	0.16	0.52
Classification Tree	0.59	0.67	0.49	0.55
Logistic regression	0.52	0.59	0.41	0.52
Neural Network	0.50	0.60	0.36	0.53
SVM	0.59	1	0	0.5

Raw_SS				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.44	0	0.98	0.57
Random Forest	0.59	0.69	0.47	0.65
kNN	0.49	0.73	0.21	0.57
Classification Tree	0.54	0.55	0.53	0.55
Logistic regression	0.56	0.55	0.57	0.56
Neural Network	0.55	0.57	0.52	0.56
SVM	0.54	1	0	0.49

Filtered_BS				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.77	0.78	0.74	0.85
Random Forest	0.73	0.80	0.62	0.79
kNN	0.72	0.87	0.51	0.76
Classification Tree	0.56	0.63	0.46	0.50
Logistic regression	0.76	0.80	0.69	0.81
Neural Network	0.74	0.80	0.65	0.81
SVM	0.6	0.98	0.09	0.69

Filtered_HS				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.75	0.73	0.78	0.83
Random Forest	0.72	0.88	0.49	0.83
kNN	0.66	0.81	0.45	0.70
Classification Tree	0.64	0.69	0.56	0.62
Logistic regression	0.79	0.86	0.69	0.81
Neural Network	0.77	0.84	0.67	0.81
SVM	0.70	0.73	0.65	0.74

Filtered_SS				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.72	0.71	0.73	0.80
Random Forest	0.69	0.76	0.60	0.81
kNN	0.61	0.63	0.60	0.66
Classification Tree	0.57	0.61	0.51	0.57
Logistic regression	0.70	0.71	0.68	0.77
Neural Network	0.67	0.69	0.65	0.76
SVM	0.60	0.88	0.26	0.68

On test data				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.88	1	0.72	0.91
Random Forest	0.6	0.71	0.45	0.72
kNN	0.68	0.85	0.45	0.75
Classification Tree	0.48	0.5	0.45	0.44
Logistic regression	0.88	1	0.72	0.95
Neural Network	0.84	1	0.63	0.95
SVM	0.52	0.92	0	0.69

On test data				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.85	0.85	0.85	0.94
Random Forest	0.76	0.9	0.57	0.90
kNN	0.67	0.95	0.28	0.73
Classification Tree	0.67	0.7	0.64	0.70
Logistic regression	0.91	0.9	0.92	0.94
Neural Network	0.91	0.9	0.92	0.94
SVM	0.61	0.75	0.42	0.76

On test data				
Classification method	CA	Sens	Spec	AUC
Naïve Bayes	0.69	0.69	0.7	0.72
Random Forest	0.67	0.69	0.65	0.71
kNN	0.58	0.56	0.6	0.63
Classification Tree	0.62	0.56	0.7	0.61
Logistic regression	0.60	0.65	0.55	0.65
Neural Network	0.58	0.65	0.5	0.67
SVM	0.58	0.86	0.25	0.58

The ROC curves for biopsy, stool, and swab datasets are shown in Figure 44, Figure 45, and Figure 46, respectively. The different classifiers are shown in different colors. The classifier that is closer to the upper left corner of the graph is better, as the area under the curve is higher. The naïve Bayes is the best performing classifier.

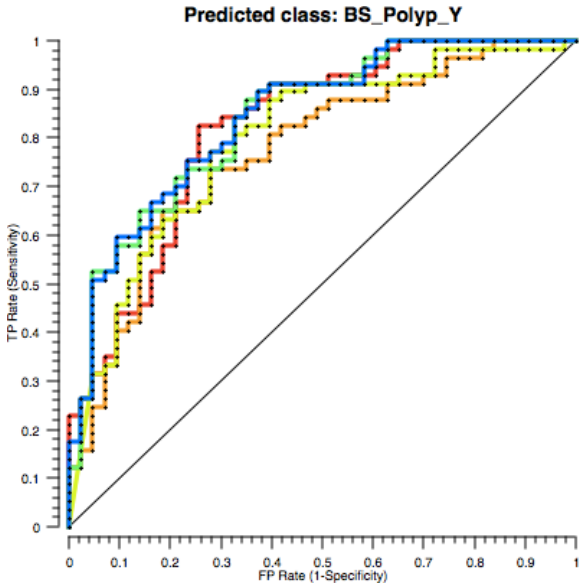


Figure 44 The polyp biopsy dataset ROC curve for the polyp-Y and polyp-N using five classifiers. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The straight line represents the null model. The x-axis is the false positive rate, and the y-axis is the true positive rate. The best performing classifier is Naïve Bayes with AUC= 0.85.

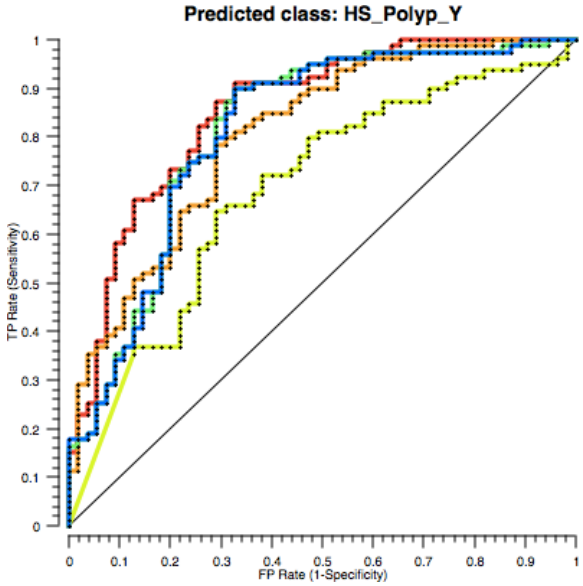


Figure 45 The ROC curve of stool dataset for the polyp-Y and polyp-N using five classifiers. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The straight line represents the null model. The x-axis is the false positive rate, and the y-axis is the true positive rate. The best performing classifiers are Naïve Bayes and random forest with AUC= 0.83.

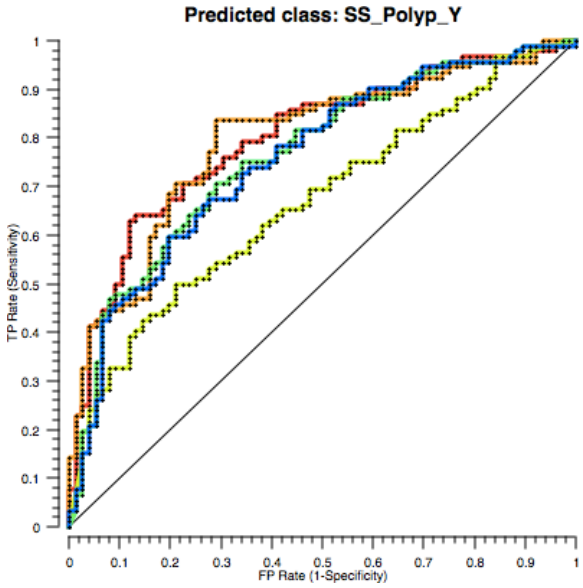


Figure 46 The ROC curve polyp rectal swab dataset for the polyp-Y and polyp-N using five classifiers. Naïve Bayes (red), Random Forest (orange), kNN (light green), Logistic regression (green), Neural Network (blue). The straight line represents the null model. The x-axis is the false positive rate, and the y-axis is the true positive rate. The best performing classifiers are the random forest with AUC=0.81 and Naïve Bayes with AUC= 0.80.

4.3.7 MB01 polyp dataset predictions

Using the prediction pipeline described in method section (Figure 19), we tested the prediction power of constructed classifiers from the UPARSE method. The separate test datasets were produced using Weka dataset generator (Weka V. 3.8.0) and used for analyzing the prediction power of each classifier.

Biopsy prediction results: From a total of 125 biopsy samples, 90% (n=112) were used for the training set and 10% (n=13) for test dataset. The training set was used to make classifiers and the resulting trained classifier was applied to the test dataset to predict classes. From 13 samples, eight of them (61.5%) were predicted correctly as showed in Table 23.

Table 23 Polyp prediction using the biopsy dataset.
Correctly classified biopsy test samples are 61.5% of the samples.

Sample	Actual Class	Predicted class using Naïve Bayes classifier
BS_9	BS_Polyp_Y	BS_Polyp_Y
BS_419	BS_Polyp_Y	BS_Polyp_Y
BS_353	BS_Polyp_Y	BS_Polyp_N
BS_431	BS_Polyp_Y	BS_Polyp_N
BS_436	BS_Polyp_Y	BS_Polyp_Y
BS_405	BS_Polyp_Y	BS_Polyp_Y
BS_351	BS_Polyp_Y	BS_Polyp_Y
BS_389	BS_Polyp_Y	BS_Polyp_Y
BS_57	BS_Polyp_N	BS_Polyp_Y
BS_517	BS_Polyp_N	BS_Polyp_Y
BS_330	BS_Polyp_N	BS_Polyp_N
BS_344	BS_Polyp_N	BS_Polyp_Y
BS_421	BS_Polyp_N	BS_Polyp_N

Stool samples prediction results: From the total of 168 stool samples, 90% (n=151) were used for training and 10% (n=17) for test dataset. From 17 predictions, 12 were correct (82.3%) as showed in Table 24.

Table 24 Polyp prediction using the stool dataset.
Correctly classified stool test samples are 82.3% of the samples.

Sample	Actual Class	Predicted class using Naïve Bayes classifier
HS_458	HS_Polyp_N	HS_Polyp_N
HS_20	HS_Polyp_N	HS_Polyp_N
HS_396	HS_Polyp_N	HS_Polyp_N
HS_370	HS_Polyp_N	HS_Polyp_N
HS_399	HS_Polyp_N	HS_Polyp_N
HS_511	HS_Polyp_N	HS_Polyp_N
HS_2	HS_Polyp_N	HS_Polyp_Y
HS_314	HS_Polyp_Y	HS_Polyp_Y
HS_62	HS_Polyp_Y	HS_Polyp_Y
HS_391	HS_Polyp_Y	HS_Polyp_Y
HS_303	HS_Polyp_Y	HS_Polyp_Y
HS_311	HS_Polyp_Y	HS_Polyp_Y
HS_318	HS_Polyp_Y	HS_Polyp_Y
HS_518	HS_Polyp_Y	HS_Polyp_Y
HS_449	HS_Polyp_Y	HS_Polyp_Y
HS_6	HS_Polyp_Y	HS_Polyp_N
HS_23	HS_Polyp_Y	HS_Polyp_N

Rectal swab samples prediction results: From the total of 211 rectal swab samples, 90% (n=189) were used for training set to constructed classifiers. Predictions are

performed on the 10% (n=22) test set. Class predictions were correct for 18/22 of the samples (81.8%) as showed in Table 25.

Table 25 Polyp prediction using the rectal swab dataset.
Correctly classified rectal swab test samples are 81.8% of the samples.

Sample	Actual Class	Predicted class using Naïve Bayes classifier
SS_3	SS_Polyp_N	SS_Polyp_N
SS_31	SS_Polyp_N	SS_Polyp_N
SS_469	SS_Polyp_N	SS_Polyp_N
SS_429	SS_Polyp_N	SS_Polyp_N
SS_24	SS_Polyp_N	SS_Polyp_N
SS_5	SS_Polyp_N	SS_Polyp_N
SS_302	SS_Polyp_N	SS_Polyp_Y
SS_25	SS_Polyp_N	SS_Polyp_N
SS_457	SS_Polyp_N	SS_Polyp_N
SS_57	SS_Polyp_N	SS_Polyp_N
SS_51	SS_Polyp_Y	SS_Polyp_Y
SS_347	SS_Polyp_Y	SS_Polyp_N
SS_516	SS_Polyp_Y	SS_Polyp_Y
SS_352	SS_Polyp_Y	SS_Polyp_Y
SS_34	SS_Polyp_Y	SS_Polyp_Y
SS_420	SS_Polyp_Y	SS_Polyp_Y
SS_362	SS_Polyp_Y	SS_Polyp_N
SS_16	SS_Polyp_Y	SS_Polyp_Y
SS_328	SS_Polyp_Y	SS_Polyp_N
SS_8	SS_Polyp_Y	SS_Polyp_Y
SS_53	SS_Polyp_Y	SS_Polyp_Y
SS_356	SS_Polyp_Y	SS_Polyp_Y

Based on these prediction results, the prediction for stool and swab samples have a higher accuracy than biopsy samples. This could be due to the larger sample size of these two datasets which results in a larger training dataset. It should be noted that with respect

to the noninvasive diagnostic aims, the swab and stool samples are preferred to biopsy as the former are noninvasive.

5 DISCUSSION

5.1 OTU clustering methods

As noted above, several algorithms have been developed for assigning 16S rRNA gene sequences to OTUs. However, these different clustering methods lead to different biodiversity estimates (Bachy et al., 2013) and, therefore, it is essential to understand the advantages and disadvantages of these methods to be able to decide which one to use for a given dataset. These methods are different in many aspects including user-friendliness, accuracy, memory requirement, and speed. Thus, choosing an appropriate clustering method can be challenging for researchers. Critical limitations of the OTU-based methods are that clustering algorithms are computationally intensive, relatively slow, and may need a considerable amount of memory (Schloss & Handelsman, 2005; Schloss et al., 2009; Sun et al., 2009).

Most *de novo* clustering algorithms (i.e., without using reference sequences) utilize either a hierarchical or greedy heuristic approach to generate clusters (Sun et al., 2012). In the hierarchical approach, a distance matrix is first calculated by measuring the difference between each pair of sequences and then the standard hierarchical clustering is employed to define OTUs at a particular level of sequence similarity (Chen et al., 2013). However, greedy heuristic algorithms perform fewer pairwise comparisons in order to estimate optimal clustering parameters that will improve computational efficiency (Sun et al., 2012). The hierarchical clustering methods may not be suitable for large sequencing datasets due

to their intrinsic computational complexity. Thus, greedy heuristic algorithms have been developed which can significantly reduce the time and space complexity (Chen et al., 2013; Li & Godzik, 2006; Sun et al., 2009; Edgar, 2010; Edgar, 2013). There is a trade-off between complexity and accuracy of the hierarchical and heuristic clustering methods. The heuristic clustering algorithms have a lower complexity at the cost of less biological accuracy (Cai & Sun, 2011; Ghodsi et al., 2011).

Some of the available hierarchical clustering methods are the nearest neighbor, furthest neighbor, weighted neighbor, and average neighbor (UPGMA) algorithms (Legendre P. & Legendre L., 1998). Among these, the average neighbor algorithm (i.e., UPGMA) is reported to perform better than the rest (Schloss & Westcott, 2011).

In the present study, we compared three *de novo* OTU clustering methods from both hierarchical and heuristic approaches that are commonly used to assign sequences into OTUs based on the similarity of sequences. Specifically, we used the greedy heuristic algorithms of UCLUST and UPARSE in addition to the hierarchical clustering algorithm UPGMA. The UPGMA algorithm takes a multiple sequence alignment as input, and after making a distance matrix of all pairwise comparisons of sequences, it starts generating clusters (Schloss et al., 2009). Before putting the alignments into the UPGMA pipeline in Mothur, we performed multiple sequence alignments with default settings in MAFFT v7.150b (Katoh and Standley, 2013) as Bachy et al., 2013 and Flynn et al., 2015. For UCLUST, the sequences were first sorted by length and then serially clustered which means the longest read in the file was the first OTU. For UPARSE, sequences were sorted based on abundance, and they were used to assign OTUs in the order of decreasing

abundance (Edgar, 2013). Among these three approaches, UPARSE was much faster than the other two methods when applied to the datasets we have analyzed. Alpha and beta diversity results of these three methods were different with respect to the diversity values due to the difference in the number of detected OTUs by each method. However, the pattern of diversity was the same, which indicates that all methods detected the same trends of diversity in the datasets. In the classification methods, classifiers that were produced from the UPARSE method performed better than UCLUST and UPGMA with respect to the accuracy, sensitivity, specificity, and area under ROC curve.

In a study done by Edgar (Edgar, 2013), the results of OTU selection using UPARSE, Mothur (UPGMA), and QIIME (UCLUST) on artificial ('16S mock') communities of known composition were compared. In all 16S mock datasets, the vast majority of UPARSE OTUs were classified as identical to the input biological sequences with less than 1% errors. On the other hand, from 41 to 71% of the Mothur OTUs and 23 to 67% of the QIIME OTUs were chimeric. In his analysis, QIIME detected more OTUs than UPARSE and Mothur, in addition to returning a large number of chimeric OTUs. The number of OTUs plus contaminants detected by UPARSE corresponded with real species and contaminants in the mock data. The strongest correlation between the number of reads and the number of OTUs were reported for QIIME, which means that the number of OTUs produced by QIIME tends to increase with respect to the number of reads, albeit because of artifactual OTUs as this pipeline lacks filtering (Edgar, 2013).

Flynn and colleagues tested Mothur, UCLUST, and UPARSE on a mock community and a natural community of zooplankton species. They reported that Mothur

gave them comparable results to UCLUST regarding OTU number and precision. However, Mothur required more time and computational resources than UCLUST and UPARSE. In their research, UPARSE showed the highest precision and the number of OTUs detected by this method was closest to the species number they used as the input. They recommended UPARSE as the method of choice for clustering among these three algorithms (Flynn et al., 2015). Although similar work from another group (Sun et al., 2012) had shown that hierarchical clustering produced better results for bacterial 16S sequences, they also reported that greedy heuristic clustering has a comparable accuracy to hierarchical clustering.

In this study, among UPARSE, UPGMA, and UCLUST, we suggest that UPARSE is our choice as preprocessing and clustering took significantly lower time and the classifiers that were produced by UPARSE OTUs performed better in the classifiers.

5.2 Feature selection to improve classification

We introduced feature selection as an optional step for classification and this was performed to find a combination of feature subsets that would lead to better classifiers. We found that feature selection resulted in significant improvement in the classification of the above datasets as there were a large number of features (OTUs) in a microbiome abundance table that far outnumbered the samples. In the field of machine learning, it is recommended that the number of samples be three times the number of features for accurate classification. Therefore, by choosing an informative subset of features, we significantly improved the classifiers' performance. Feature screening also improved prediction accuracy and led to generating more easily interpretable models (Knights et al., 2011). Many approaches have

been introduced for feature selection including filter methods, wrapper methods, and embedded methods. Filter approaches are the most straightforward method in which features are selected on the basis of statistical properties and are performed before classification. A univariate test such as the *t-test* or a multivariate test like linear classifier test is conducted to select the features that have a score above a predefined threshold (e.g., 0.05 significance level). Filter methods have a number of advantages including low computational complexity and ease of implementation (Knights et al., 2011), while wrapper methods are computationally intensive. Like filter methods, the wrapper methods treat the classifier as a black box. However, this approach uses a classifier to select a subset of features. As the classifier needs to examine all the feature subsets in order to find the one with the lowest validation error, this method is considered computationally intensive (Knights et al., 2011). Embedded approaches are an integral part of the machine learning process as these methods run an integrated search over the joint space of model parameters and feature subsets. The advantage of this approach is that it can look for globally optimal parameters (Knights et al., 2011).

In our study, we used independent nonparametric statistical tools to find significant features between binary groups in order to select OTUs that are associated with a shift from normal state to disease state to decrease the complexity of the analysis.

MetaStats (White et al., 2009) was used to find OTUs that have significantly different mean proportion and variance among groups using Kruskal-Wallis, which is a non-parametric *t-test*. MetaStats combines statistical analysis with biomarker discovery

based on repeated Kruskal-Wallis and Fisher's exact tests on random permutations (Segata et al., 2011).

The Kruskal-Wallis test is a non-parametric statistical approach that tests whether samples originate from the same distribution. This test does not require the data to be normal but rather uses the rank of the data values instead of the actual data values for the analysis. OTUs that have a significantly different mean rank between groups will be distinguished by this test.

Another approach, LEfSe (Segata et al., 2011), starts with a non-parametric factorial Kruskal-Wallis sum-rank test to find features that have a significant difference in abundance between classes and then performs a series of pairwise tests among subclasses using unpaired Wilcoxon rank-sum test to find significant biological features. It finally applies Linear Discriminant Analysis to measure the effect size of each differentially abundant feature.

The final tool that we used was Indicator which uses relative abundance and the relative frequency of occurrence to detect OTUs that are distinctive features (OTUs) of each of the groups under study. Using these four approaches, we identified significant features between groups and used them for training classifiers. We demonstrated that using statistically significant OTUs instead of all OTUs as classification features considerably improves classification accuracy, sensitivity, and specificity.

5.3 Classification and data mining in microbiome studies

Microbiome analyses are generally restricted to measuring taxon relative abundancies, analyzing alpha and beta diversity, exploring beta diversity patterns using

unsupervised learning techniques such as clustering and PCoA, and performing classical hypothesis testing. However, these methods are not able to classify unknown or unlabeled data or to extract noticeable features from highly complex or sparse datasets (Knight et al., 2011). Supervised machine learning methods are useful for finding patterns in highly complex datasets like human microbiota surveys and help in finding predictive features to find the class for unlabeled data (Wisittipanit 2015). OTU abundance tables produced from the sequencing data and upstream bioinformatics analysis pipelines can be used as input training data for supervised machine learning methods to develop predictive models. In microbiome studies, the training data consists of the relative abundance of OTUs and a categorical variable that denotes the correct classification of that data (e.g., cancer and healthy disease states). The purpose of supervised classification is to derive a model from the training data with known classes and use it for assigning the correct class or category labels to new samples with unknown classes and identifying the features (OTUs) that can discriminate between classes (Knights et al., 2011). The classification model should neither be very general, as it would not be able to incorporate subtle but critical information (underfitting) nor too complicated, as it would be accurate for that particular dataset and not useful for a novel dataset (overfitting) (Knights et al., 2011).

The study by Beck and Foster is one of the examples of using machine learning methods to make models of classification for a disease (Beck & Foster, 2014). They used eight different classifiers on bacterial vaginosis (BV) sequencing datasets. They found that classification models which were produced using genetic programming, random forests, and logistic regression can classify microbial communities into BV categories with an

accuracy of 80 to 90%. They tried to deconstruct classifier models to find out which features are essential for the accuracy but they observed different features for each classifier (Beck & Foster, 2014). Therefore, they designed another study to determine the critical features by adding the features sequentially to the models. In this study, they only used random forests and logistic regression classifiers. They showed that the models that are generated by logistic regression and random forests approaches perform the same and the main features that were detected were very similar. They concluded that only a few features were required for obtaining high accuracy and that most of the features were redundant (Beck & Foster, 2015).

We undertook a comparison of eight classification methods -- four feature selection approaches and four accuracy metrics -- for three different datasets from different sequencing platforms. We focused on supervised classification methods, as unsupervised methods like clustering are designed to reveal the structure of the data to provide visual summaries and to help quality control, but they are not suitable for predictions and assigning naïve data to a specific class (Dupuy et al., 2007; Simon et al., 2003). We first performed classification with the complete relative abundance OTU table using all OTUs and then with relative abundance tables of just the significant OTUs. For all three datasets and all three OTU selection methods, the accuracy of classification, sensitivity, and specificity of classification as well as the area under the curve were improved by using the significant features demonstrating that most of the OTUs were redundant and removing them from the beginning improved the model performance. Furthermore, we could

conclude that decreasing the number of OTUs reduces the complexity of classification, leading to better-performing models.

5.4 Bacterial shifts in CRC and Adenoma

Intestinal microbiota starts to develop before birth, and it will continue to evolve and increase the diversity by 3-5 years of age. Many factors are contributing to form the microbial structure such as genetics, delivery mode, diet, however, when it settles, the composition will not change dramatically unless an intervention like infection or antibiotics occurs. Any microbial shift will enhance the chance of disease development (Rodriguez et al., 2015). There are two bacterial compartments in the colon, the luminal and the mucosal. The luminal microbiota is thought to be transient, changes with diet, and is not representative of the localized epithelial and cryptal microbiota (Savage, 1977). On the other hand, the mucosal microbiota adheres to the surface-associated polysaccharide matrices of the colon and are resistant to colonic movement. Our findings show that the microbial structure of the intestinal lumen differs from the biopsies as reported in a number of previous studies (Chen et al., 2012; Mira-Pascual et al., 2015).

Comparing detected bacterial taxa from different studies is challenging as they differ in important aspects such as sample collection and storage, DNA extraction methods, sequencing technology, chosen 16S variable region, and sequence analysis pipelines. Additionally, the 16S rRNA reference databases are different in many cases and these differences need to be considered when comparing the results of these studies. All of these aspects can affect the bacterial species that are identified in these various studies. In addition to this lack of standardization, the nature of the gut microbiome is inherently

dynamic and can differ among people based on their genetic background, host immune system, diet, age, geography, and use of antibiotics and other medications (Goodrich et al., 2014; Hooper et al., 2012; Turnbaugh et al., 2008; Yatsunenko et al., 2012). Thus, different studies with different participants and methods may produce dissimilar results. Additionally, it is not clear whether changes in a few bacteria lead to altering health status or that the disease state changes a specific taxonomic group of bacteria or that changes the whole microbiota (i.e., dysbiosis) to induce disease. Therefore, studying the microbial changes in different health status becomes complex and sometimes confounded.

Although bacterial dysbiosis has been reported in virtually all CRC/ adenoma microbiome studies as well as our research, the microbial features that have been found to differ significantly are not the same, or in some cases, are contradictory. Even for taxa that were shown to differ in the disease state, the direction of the shift is different in various studies. In addition, the bacterial abundance changes in different states of the disease have been reported to be different as reported in our study.

5.4.1 Comparing alpha diversity results with other CRC/ adenoma studies

A summary of alpha diversity results from this study and seventeen other CRC and adenoma studies along those phyla significantly changed are summarized in Table 26.

Six of these seventeen studies reported higher alpha diversity in adenoma or cancer groups, four studies observed lower diversity in the adenoma or cancer group, and three studies indicated no significant change in the diversity of the cancer or adenoma group compared to healthy control subjects (Shen et al., 2010; Sanapareddy et al., 2012; McCoy et al., 2013; Mira-Pascual et al., 2015; Chen et al., 2013; Ahn et al., 2013; Goedert et al.,

2015; Hale et al., 2017). In our polyp study, alpha diversity was lower in biopsies from people with a polyp compared to those without a polyp as shown in Table 26. This lack of agreement among different studies could be attributed to the variability in study designs and technical differences or could be due to confounding clinical factors such as genetic background, diet, lifestyle, and medications. Additionally, the lack of internal standards for microbiome studies could contribute to variability among these various studies.

5.4.2 Comparing the changes of bacterial communities' composition at phylum level

The most commonly reported significant differences of microbe abundancies among healthy control, adenoma, and CRC samples are for Firmicutes, Bacteroidetes, and Proteobacteria phyla. Therefore, we compared abundancies of these three phyla in our polyp dataset and in seventeen other cohorts of patients described in the literature. Firmicutes and Bacteroidetes are the predominant phyla in the healthy human gut (Jandhyala et al., 2015). Reportedly, Firmicutes enhance energy harvest from the diet, while Bacteroidetes are involved in interactions with the mucosa (Costello et al., 2010; Ley et al., 2006; Turnbaugh et al., 2006; Joly et al., 2010). Proteobacteria directly interact with intestinal cells through bacterial secretion systems T2SS or T3SS (Beeckman & Vanrompay, 2010; Brown & Finlay, 2010).

Some studies have shown increased abundance of Firmicutes in adenoma/ CRC samples while other studies reported relative depletion of this phylum (Sanapareddy et al., 2012; Brim et al., 2013; Marchesi et al., 2011; Kostic et al., 2012; Wu et al., 2013; Ahn et al., 2013; Hale et al., 2017; Goedert et al., 2015). However, one study reported no

significant changes of abundance of this phylum in adenoma biopsies (Shen et al., 2010) as seen in Table 26.

Reports of changes in Bacteroidetes abundance in disease states are also confounded. A number of reports described an increase in abundance of Bacteroidetes in patients with adenoma/ CRC while others reported a decrease in abundance of this phylum (Marchesi et al., 2011; Sanapareddy et al., 2012; Wu et al., 2013; Ahn et al., 2013; Mira-Pascual et al., 2015; Goedert et al., 2015; Shen et al., 2010; Kostic et al., 2012; McCoy et al., 2013; Brim et al., 2013). These studies are summarized in Table 26. In our polyp dataset, Bacteroidetes and Firmicutes phyla were significantly more abundant in the polyp-Y samples (Figure 47, Figure 48, and Figure 49).

Table 26 Comparison of the alpha diversity and the change of taxa abundance at phylum level in different adenoma and CRC studies.
There is no consensus among various studies in the alpha diversity and bacterial abundance at phylum level for Bacteroidetes, Firmicutes, and Proteobacteria. In most of the studies, alpha diversity increased in the adenoma and cancer group. Proteobacteria has been reported to increase in the adenoma/ CRC state in most of the studies. Ad: Adenoma; CRC: colorectal cancer; H: Healthy; N/A: not available; N/S: not significant.

Study	Samples	Disease state	Alpha diversity	Firmicutes	Bacteroidetes	Proteobacteria
Shen et al. 2010	Mucosal biopsy	Adenoma	Ad>H	N/S	Ad<H	Ad>H
Marchesi et al. 2011	Tumor/adjacent normal tissue	CRC	N/A	CRC<H	CRC>H	N/A
Chen et al. 2012	rectal swabs, fecal samples, tumor/ matching normal tissue	CRC	Tissue: Tumor< normal	No change	Tumor> normal	Tumor< normal
			Swab:	CRC<H	No change	CRC<H
			Stool	CRC<H	No change	CRC>H
Kostic et al. 2012	Tumor/matching normal	CRC	N/A	Tumor< normal	Tumor< normal	N/A
Sanapareddy et al. 2012	Rectal mucosa biopsy	Adenoma	Ad>H	Ad>H	Ad>H	Ad>H
Wu et al. 2013	Fecal	CRC	N/S	CRC<H	CRC>H	N/S
McCoy et al. 2013	Rectal mucosa biopsy	Adenoma	CRC>H	N/A	CRC<H	N/A
Brim et al. 2013	Fecal	Adenoma	N/A	Ad>H	Ad<H	Ad>H
Ahn et al. 2013	Fecal	CRC	CRC<H	CRC<H	CRC>H	N/A
Zackular et al. 2014	Fecal	CRC & adenoma	N/A	N/A	N/A	N/A
Mira-Pascual et al. 2015	Fecal and biopsy	CRC & adenoma	CRC>Ad>H	N/A	Biopsy: CRC & Ad>H	N/A
Goedert et al. 2015	Fecal	CRC & adenoma	No change	Ad<H	Ad>H	Ad>H
Thomas et al., 2016	Biopsy	CRC	CRC>H	N/A	CRC>H	CRC<H
Xu and Jiang 2017	Biopsy	CRC & adenoma	Ad<H (N/S) CRC>H (N/S)	Ad<H (N/S) CRC>H	N/A	CRC<H Ad>H
Gao et al. 2017	Tumor/matching normal	CRC	No significant change	Tumor< normal	Tumor> normal	Tumor> normal
Yoon et al. 2017	Biopsy	CRC & adenoma	CRC<H Ad<H	CRC<H Ad>H	CRC<H Ad>H	CRC>H Ad<H
Hale et al. 2017	Fecal	Adenoma	No change	Ad<H	Ad>H	N/A
Our polyp dataset 2016	Fecal/ rectal swab/ biopsy	Polyp	Biopsy: Polyp-Y<polyp-N	polyp-Y>Polyp-N	polyp-Y>Polyp-N	polyp-Y<Polyp-N
			Swab: No change	polyp-Y>Polyp-N	polyp-Y>Polyp-N	polyp-Y>Polyp-N

Study	Samples	Disease state	Alpha diversity	Firmicutes	Bacteroidetes	Proteobacteria
			stool: No change	polyp-Y>Polyp-N	polyp-Y>Polyp-N	polyp-Y>Polyp-N

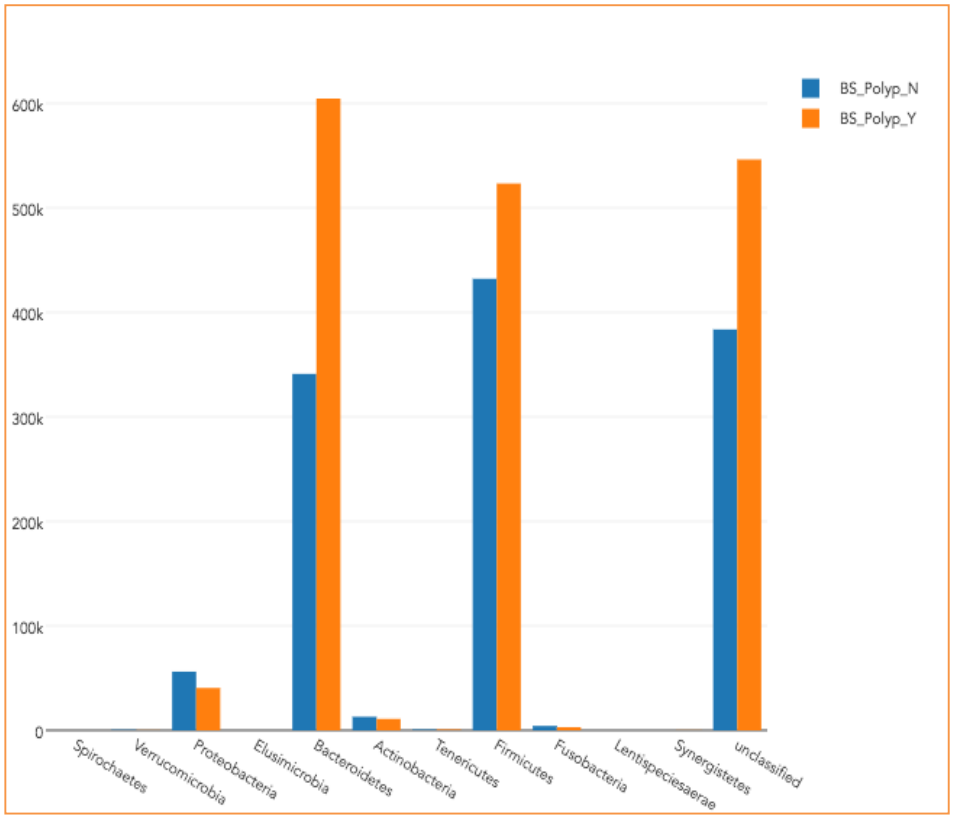


Figure 47 Bacterial changes at phylum level in the polyp biopsy (BS) samples.
The dominant phyla of the colon, Bacteroidetes, and Firmicutes are increased in the polyp-Y group compared to the polyp-N group.

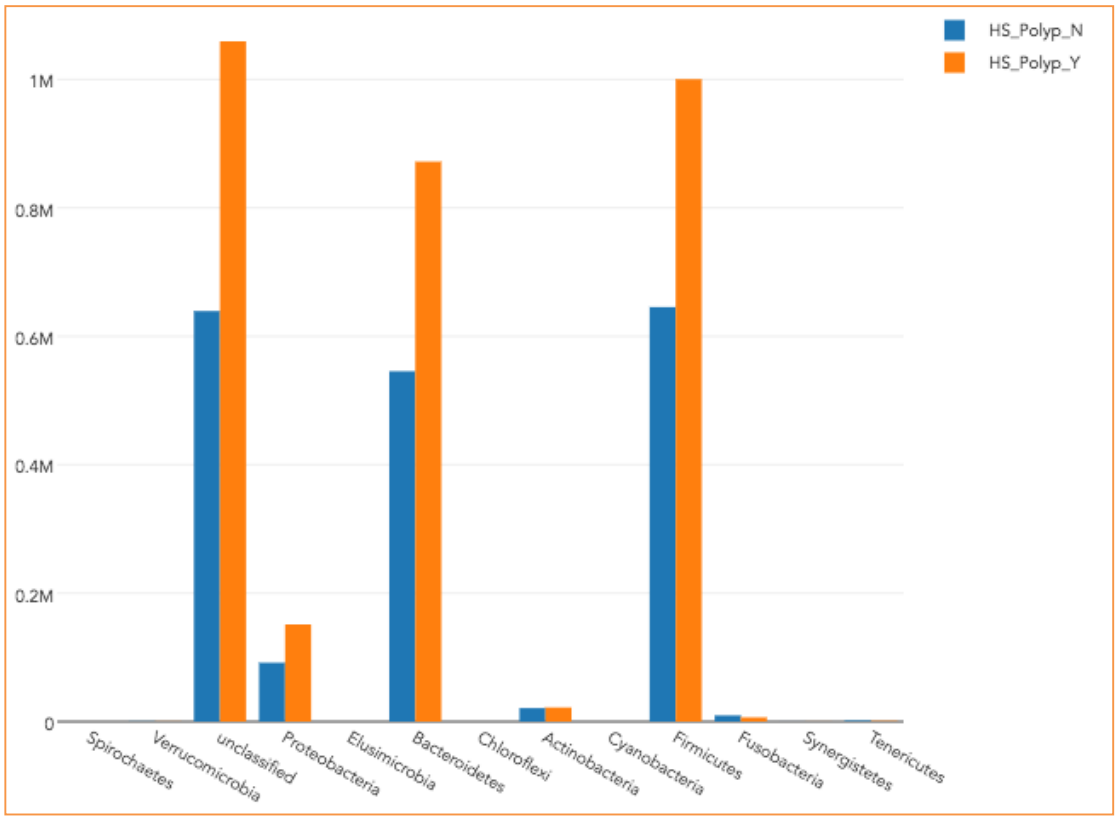


Figure 48 Bacterial changes at phylum level in the polyp stool (HS) samples.
The most dominant phyla of the colon, Bacteroidetes, Firmicutes, and Proteobacteria are increased in the polyp-Y group compared to the polyp-N group.

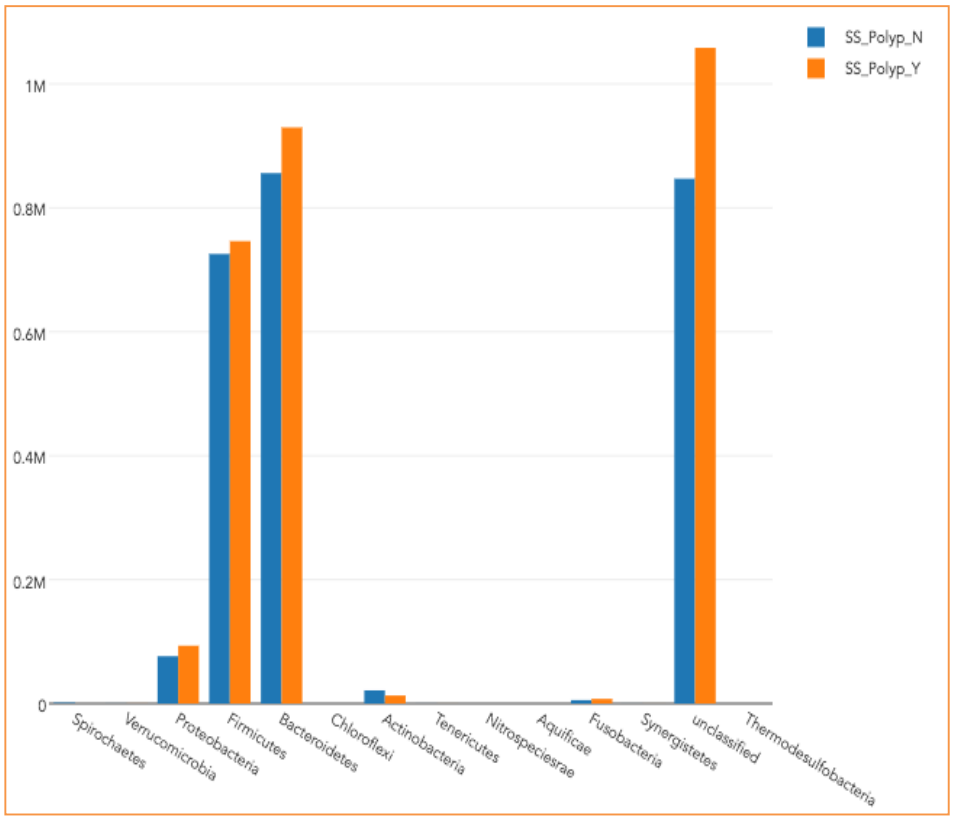


Figure 49 Bacterial changes at phylum level in the polyp swab (SS) samples.
The most dominant phyla of the colon, Bacteroidetes, Firmicutes, and also Proteobacteria are increased in the polyp-Y group compared to the polyp-N group.

Among the seventeen studies of adenoma/ CRC listed in Table 26, eight datasets contained information on Proteobacteria. Speaking generally, observations on Proteobacteria abundance in the adenoma studies are somewhat more consistent than that for Bacteroidetes. In five out of six adenoma studies which reported Proteobacteria abundance, this phylum was represented at higher levels in the adenoma state (Shen et al., 2010; Sanapareddy et al., 2012; Goedert et al., 2015; Brim et al., 2013, Xu & Jiang, 2017). Only one, relatively small study reported lower Proteobacteria abundance in the adenoma state (Yoon et al., 2017). Among six CRC studies reporting the Proteobacteria abundancies

there was no consensus. Two studies detected higher Proteobacteria abundance with CRC, and two datasets reported lower abundance with CRC (Thomas et al., 2016; Xu & Jiang, 2017; Gao et al., 2017; Yoon et al., 2017). In one study that analyzed three types of samples -- tumor, swab, and stool -- an increase of Proteobacteria abundance in the stool of the CRC group and a decrease of this phylum in both the tumor and swab samples were reported (Chen et al., 2012).

In our own polyp dataset, analysis of both rectal swabs and stool samples revealed an increase in Proteobacteria abundance. However, in biopsies of the polyps, the abundance of Proteobacteria was paradoxically lower compared to the normal colon biopsies (Figure 47).

There is no consensus on the Firmicutes and Bacteroidetes changes with CRC when compared to the healthy state. As it is shown in Table 26, previous studies do not agree with each other. As mentioned before, one of the main reasons for this discrepancy could be the lack of standards for microbiome analysis or some unaccounted intrinsic differences between the microbiome of these cohorts. Another factor could be the sample size, as some of these studies recruited a low number of subjects. This observation is also true for other taxonomic levels. For example, in a study performed on CRC and adenoma samples, it has been shown that there are many genera, like *Blautia* and *Prevotella*, that were absent in healthy control subjects and present with polyps or tumors (Mira-Pascual et al., 2015). However, in our biopsy samples, both genera were present in both groups, and they were lower in polyp-Y compared to polyp-N, as shown in Figure 41.

5.4.3 Comparing the changes of bacterial communities' composition at the genus level

In previous CRC/ adenoma studies, significant changes in abundance between disease and healthy control groups were reported for some genus-level taxa. A few examples of this kind are *Fusobacterium*, *Bacteroides*, *Blautia*, *Bifidobacterium*, *Roseburia*, and *Faecalobacterium* (Marchesi et al., 2011; Chen et al., 2012; Kostic et al., 2012; Wu et al., 2013; Ahn et al., 2013; Mira-Pascual et al., 2015; Zackular et al., 2014; Shen et al., 2010; Brim et al., 2013; Goedert et al., 2015; Nugent et al., 2014; Chen et al., 2013). Table 27 shows the direction of change in abundance for these taxa in the studies cited above and in the MBO1 polyp dataset.

Table 27 Comparison of the frequently reported genera between our polyp dataset and seventeen previous adenoma/ CRC studies.
For most of these taxa, no consensus conclusion of enrichment or depletion can be deduced from these studies. *Fusobacterium* showed enrichment in all CRC groups, however, in adenoma, it showed mostly depletion than enrichment in the disease state. *Bifidobacterium* reported to significantly decrease in the CRC and adenoma groups in two studies. However, other studies have not reported these taxa as a significant feature. Ad: adenoma; CRC: colorectal cancer; H: healthy; N/A: not available; N/S: not significant.

Study	Samples	Disease state	*Fusobacterium*	*Bacteroides*	*Blautia*	*Bifidobacterium*	*Roseburia*	*faecalobacterium*
Shen et al. 2010	Mucosal biopsy	Adenoma	N/A	Ad<H	N/A	N/A	N/A	Ad>H
Marchesi et al. 2011	Tumor/adjacent normal tissue	CRC	Tumor> normal	N/A	N/A	N/A	Tumor > normal	CRC>H
Chen et al. 2012	tumor/ matching normal tissue Intestinal lumen, mucosa (rectal swabs), fecal samples,	CRC	Tissue: N/A	Tumor> normal	Tumor < normal	N/A	Tumor < normal	CRC>H
			Swab: CRC>H	N/A	CRC< H	CRC<H	N/A	CRC<H
			Stool: N/A	N/A	N/A	N/A	N/A	N/A
Kostic et al. 2012	Tumor/matching normal	CRC	Tumor> normal	N/A	N/A	N/A	N/A	N/A
Sanapareddy et al. 2012	Rectal mucosa biopsy	Adenoma	N/A	N/A	N/A	N/A	N/A	N/A
Wu et al. 2013	Fecal	CRC	CRC>H	CRC>H	CRC> H	N/A	CRC< H	CRC<H
McCoy et al. 2013	Rectal mucosa biopsy	Adenoma	Ad>H	N/A	N/A	N/A	N/A	N/A
Brim et al. 2013	Fecal	Adenoma	N/A	Ad<H	N/A	N/A	No change	Ad>H
Ahn et al. 2013	Fecal	CRC	CRC>H	N/A	N/A	N/A	N/A	N/A
Zackular et al. 2014	Fecal	CRC & adenoma	CRC>H	Ad<H CRC<H	Ad>CRC	N/A	N/A	N/A
Mira-Pascual et al. 2015	Fecal and biopsy	CRC & adenoma	Mucosal: CRC>H	Stool: CRC>H	Biopsy: Ad>H	Stool: Ad<H CRC<H	N/A	N/A

Study	Samples	Disease state	*Fusobacterium*	*Bacteroides*	*Blautia*	*Bifidobacterium*	*Roseburia*	*faecalobacterium*
					CRC>H			
Goedert et al. 2015	Fecal	CRC & Adenoma	Ad<H	Ad<H Ad>H	Ad<H Ad>H	N/A	N/A	N/A
Thomas et al., 2016	Biopsy	CRC	CRC>H	CRC>H	N/A	N/A	CRC>H	N/A
Xu and Jiang 2017	Biopsy	CRC & adenoma	CRC>H Ad<H	CRC>Ad>H	CRC<H	N/A	N/A	CRC<H
Gao et al. 2017	Tumor/matching normal	CRC	Tumor> normal	Tumor> normal	N/A	N/A	N/A	Tumor< normal
Yoon et al. 2017	Biopsy	CRC & adenoma	N/A	CRC<H Ad>H	CRC<H Ad>H	N/A	N/A	CRC<H Ad>H
Hale et al. 2017	Fecal	CRC	No enrichment in adenoma	N/A	N/A	N/A	N/A	N/A
Our polyp dataset 2016	Fecal/ swab/ biopsy	Polyp	Biopsy: polyp-Y<polyp-N (N/S)	polyp-Y> polyp-N	polyp-Y< polyp-N	No change	polyp-Y< polyp-N (N/S)	polyp-Y< polyp-N
			Swab: polyp-Y< polyp-N (N/S)	polyp-Y> polyp-N	polyp-Y> polyp-N	No change	polyp-Y> polyp-N (N/S)	polyp-Y> polyp-N
			Stool: polyp-Y< polyp-N (N/S)	polyp-Y> polyp-N	polyp-Y> polyp-N	No change	polyp-Y> polyp-N (N/S)	polyp-Y< polyp-N

***Fusobacterium*:** Ten of seventeen studies summarized in Table 27 reported an increase of *Fusobacterium* abundance in cancer patients compared to healthy control subjects (Marchesi et al., 2011; Chen et al., 2012; Kostic et al., 2012; Wu et al., 2013; Ahn

et al., 2013; Zackular et al., 2014; Mira-Pascual et al., 2015; Thomas et al., 2016; Xu & Jiang, 2017; Gao et al., 2017). Among nine adenoma studies listed in Table 27, only three reported associations of adenomatous growth with *Fusobacterium* (McCoy et al., 2013; Goedert et al., 2015; Xu & Jiang, 2017). McCoy et al., observed that *Fusobacterium* abundance was significantly higher in adenomas as compared to non-adenoma samples. In their study, a high level of *Fusobacterium* increases the chance of finding adenoma by 3.5 fold. Moreover, a positive correlation between *Fusobacterium* presence and cytokine levels was reported in the adenoma cohort. Specifically, a significant association between the levels of TNF-α, a cell signaling protein involved in systemic inflammation, and the abundance *Fusobacterium* level was reported (McCoy et al., 2013). On the other hand, the other two studies reported a lack of *Fusobacterium* enrichment in the adenoma group (Goedert et al., 2015; Xu & Jiang, 2017). Another recent study also reported no significant change in the abundance of *Fusobacterium* in the advanced and non-advanced adenoma group as compared to colonoscopically healthy control adults (Amitay et al., 2017).

In the MBO1 polyp cohort, *Fusobacterium* only showed a marginal decrease of its abundance in the biopsy, stool, and rectal swabs. It is possible that the pathways of polypogenesis are different from that of carcinogenesis and may depend on the invasiveness of the polyps which may, in turn, be reflected in observed discrepancies of association of *Fusobacterium* abundance. Additionally, the different experimental and analytical pipelines among the studies could be confounding this observation.

Kostic et al. showed that *Fusobacterium* changes the tumor immune microenvironment in a way that could induce inflammation and tumorigenesis which, in

turn, may promote adenoma and CRC (Kostic et al., 2012). They suggested that early somatic mutations might generate an optimal environment for *Fusobacterium spp*. to colonize the mucosa, while the subsequent progression of colonization could promote myeloid cell-mediated immune responses capable of activating inflammatory pathways and stimulating further progression of the tumor (Kostic et al., 2013). The products generated by *Fusobacteria*, such as formyl-methionyl-leucyl-phenylalanine and short chain fatty acids, are also reported as myeloid cell chemoattractants. In turn, the expansion of myeloid-derived immune cell types can promote tumor progression (Qian and Pollard, 2010). The abundance of *Fusobacterium* species has also been shown to be correlated with inflammatory bowel diseases (IBD), including both ulcerative colitis and Crohn's disease (Neut et al., 2002; Ohkusa et al., 2002; Strauss et al., 2011). Notably, ulcerative colitis is considered as one of the most important risk factors for colorectal cancer.

Bacteroides*: Bacteroides* species are gram-negative, anaerobic, bile-resistant, non-spore-forming butyrate-producing bacteria. Butyrate is a short chain fatty acid that has been shown to be effective in preventing inflammation through regulatory T-cells and controls proinflammatory cytokine expression (Cushing et al., 2015; Furusawa et al., 2013; Chang et al., 2014). This short chain fatty acid is also a histone deacetylase inhibitor and has a role in preventing colonic tumors and promoting normal cell proliferation, differentiation, and apoptosis. Additionally, butyrate modulates the Wnt signaling pathway involved in the development of colorectal cancer (Malcomson et al., 2015). Many *Bacteroides* species have been isolated from human stool. *Bacteroides fragilis* (*B. fragilis*) is the most common *Bacteroides* species found in clinical specimens, and it has been reported to have virulent

properties in some instances. *Bacteroides* becomes a part of the intestinal microbiota early in life as it can pass from mother to the child during the vaginal birth (Reid, 2004). *B. fragilis* is generally considered to be a beneficial bacterium, but it can be pathogenic if it escapes from the gut to other body sites such as the abdomen, brain, liver, and lungs (Wexler, 2007).

Among the studies listed in Table 27, there are reports of both decrease and increase of this genus in the CRC and adenoma cohorts with respect to the control groups. The number of studies detecting a higher abundance of *Bacteroides* in the cancer states are higher than the number of studies observing lower abundance of this genus (Chen et al., 2012; Wu et al., 2013; Zackular et al., 2014; Mira-Pascual et al., 2015; Thomas et al., 2016; Xu & Jiang et al., 2017; Gao et al., 2017; Yoon et al., 2017). However, for the adenoma cohorts, almost half detected enrichment and half reported depletion of *Bacteroides* in the adenoma state (Shen et al., 2010; Brim et al., 2013; Zackular et al., 2014; Yoon et al., 2017). There is even one study that reported both decrease and increase of *Bacteroides* taxa in the adenoma state which can be because of the differences in the subtaxa (Goedert et al., 2015). In the three datasets we studied, there was more than one OTU classified as *Bacteroides,* and there was no consistent pattern of decrease or increase in the adenoma and cancer groups compared to healthy control samples. In our polyp dataset, all detected *Bacteroides* genera were higher in the polyp-Y group than polyp-N. However, in Benchmark-1, there were 3 OTUs assigned as *Bacteroides* that either decreased or increased in adenoma compared to the healthy control group. Between the healthy control and the cancer groups, there was a decrease of *Bacteroides* in the cancer group compared

to the healthy control group. Benchmark-2 had more than one OTU assigned as *Bacteroides,* and their abundance either decreased or increased in the cancer group compared to the healthy control. As many different species and strains with potentially different functionalities can exist in the same genus, it is not that surprising that some of the *Bacteroides* taxa were decreased while others were increased in the disease state.

Enterotoxigenic Bacteroides fragilis: *Enterotoxigenic B. fragilis* (ETBF) is a virulent bacterium that produces a toxin named fragilysin or B. fragilis toxin (BFT). BFT can stimulate both inflammatory responses and cell proliferation. The inflammatory effect of BFT occurs through activating the nuclear factor kappa B which stimulates inflammatory mediators. These mediators promote inflammation which is a risk factor for CRC (Sears et al., 2009; Shiryaev et al., 2013). The proliferation effect of ETBF occurs through activating the Wnt/ β-catenin signaling pathway which increases cell proliferation (Sokol et al., 1999). In the APC minus mouse model[3], Wu et al., showed that ETBF could promote tumorigenesis and increases the chance of colon adenoma and tumor formation in mice colonized with ETBF compared to control mice (Wu et al., 2009). Tumorigenesis effect of ETBF can also occur through other pathways such as activating STAT3, inducing IL-17 cytokine, and inducing spermine oxidase-dependent reactive oxygen species (ROS) production which causes DNA damage (Tosolini et al., 2011; Goodwin et al., 2011). These studies suggest that there is a link between bacterial antigens, virulence factors, colon adenomas, and CRC.

[3] Multiple intestinal neoplasia mouse that carries a truncation mutation at codon 850 of the *Apc* gene. The *Min* mouse can develop up to 100 polyps in the small intestine in addition to colon tumors.

In Benchmark-1, *B. fragilis* abundance was higher in cancer compared to adenoma. In Benchmark-2, there was a significantly higher abundance of *B. fragilis* in the cancer group compared to the healthy control group. However, in MBO1 polyp dataset, the level of *B. fragilis* showed a reduction in the polyp-Y group. Thus, our study does not support *B. fragilis* playing a role in the polypogenesis pathway.

Blautia: *Blautia* is a butyrate-producing bacterium that belongs to the *Firmicutes* phylum and Clostridial order. Most butyrate producers in the human colon belong to the *Firmicutes* phylum and in particular the clostridial clusters IV and XIVa (Louis and Flint, 2009; Van den Abbeele et al., 2013; Vital et al., 2014). *Blautia* is known to digest complex carbohydrates, and an abundance of these bacteria is a strong indication of a healthy gut. As such, it is usually reported that the abundance of *Blautia* decreases in colorectal cancer patients as compared to the healthy control individuals (Chen et al., 2012; Goedert et al., 2015; Xu & Jiang, 2017; Yoon et al., 2017). However, there are also reports of an increased abundance of this genus in the disease state (Wu et al., 2013; Mira-Pascual et al., 2015; Goedert et al., 2015 Yoon et al., 2017).

In the Benchmark-1, the abundance of two *Blautia* OTUs was low in the CRC group when compared to the healthy control group. However, there was another *Blautia* OTU that was increased in the adenoma group compared to the healthy control. In Benchmark-2, the genus *Blautia* was significantly lower in the cancer group compared to the healthy group.

In the polyp dataset, two different OTUs were assigned as *Blautia* in the biopsy samples that were significantly lower in the polyp-Y group compared to the polyp-N group.

In stool samples and rectal swabs, there were two OTUs classified as *Blautia*, and they were higher in the polyp-Y group. Based on these observations, *Blautia* was decreased in the CRC state, but in the adenoma state either increases or decreases can be seen. Differential abundance may depend on whether the adenoma is benign or aggressive suggesting that the potential effect of strains of this genus on the polyposis or carcinogenesis pathways may be different. As we did not have additional detailed clinical information about the patients with the polyps, we could not further clarify the reason behind this pattern in different sample types.

Bifidobacterium: Members of the genus *Bifidobacterium* are gram-positive anaerobic bacteria that are part of the gastrointestinal tract, vagina, and mouth microbiota (Duranti et al., 2016). This genus is one of the main genera of *Actinobacteria* that makes up the colon microbiota in mammals. In infants, *Bifidobacterium* may constitute 95% of the fecal microbiota of breastfed babies, but after weaning and exposure to food-derived and environmental microorganisms, the relative abundance of *Bifidobacterium* is reduced. In adults, *Bifidobacterium* makes up about 3–6% of all bacteria. *Bifidobacterium* is known to produce short-chain fatty acids that decrease the gut pH, form biological barriers, and secrete anti-microbial compounds that attenuate harmful bacteria (Bottacini et al., 2016; Liao et al., 2016). Chen et al. and Mira-Pascual et al. have shown *Bifidobacteria* depletion in the cancer and adenoma patients (Chen et al., 2013; Mira-Pascual et al., 2015). However, in other studies and in our polyp dataset, there was no significant change for this genus possibly because the abundance levels were too low for detection.

Roseburia: *Roseburia* spp. are commensal bacteria that produce short-chain fatty acids, particularly butyrate which affects colonic motility and immunity maintenance and has anti-inflammatory properties. Depletion in *Roseburia* spp. abundance may affect various metabolic pathways and be associated with several diseases including irritable bowel syndrome, obesity, Type 2 diabetes, nervous system conditions, and allergies. *Roseburia* spp. could also serve as probiotics for the restoration of a beneficial microbiota (Tamanai-Shacoori et al., 2016). As *Roseburia* produces large amounts of butyrate by fermenting dietary carbohydrates, it may be critical for the control of inflammatory processes, especially in the colon (Louis et al., 2010; Louis et al., 2014; Pryde et al., 2002; Tamanai-Shacoori et al., 2016).

Two of the seventeen studies reported enrichment of *Roseburia* in CRC samples, and two other studies showed depletion of this taxon in the cancer group (Marchesi et al., 2011; Thomas et al., 2016; Chen et al., 2012; Wu et al., 2013). None of the adenoma studies we investigated showed changes in *Roseburia* taxa. In MBO1 polyp dataset, we did detect a decrease in the abundance of this taxon in biopsy samples and an increase in abundance in the swab and fecal samples of the polyp-Y group, although these changes were not statistically significant.

Faecalibacterium: *Faecalibacterium* is commonly present in the gastrointestinal tract and is recognized as a commensal bacterium. *Faecalibacterium prausnitzii* (*F. prausnitzii*) is a dominant species of the Clostridium leptum group and is one of the most abundant anaerobic bacteria in the human gut (Arumugam et al., 2011). *F. prausnitzii* plays a key role in maintaining intestinal health and providing energy to the colonocytes (Louis

& Flint, 2009). It has also been shown that *F. prausnitzii* levels were decreased in IBD patients compared with healthy control controls (Yang et al., 2008). Three of the studies listed in Table 27 found that *Faecalibacterium spp.* were increased in adenoma subjects (Shen et al., 2010; Brim et al., 2013; Yoon et al., 2017) and Marchei et al. reported an increase of this genus in CRC cases (Marchesi et al., 2011). Chen et al. reported an increase of *Faecalibacterium* in CRC tissue while they found a decrease of this taxon in the swab samples (Chen et al., 2012). Four other studies demonstrated a decrease of *Faecalibacterium* in the CRC group (Wu et al., 2013; Xu & Jiang, 2017; Gao et al., 2017; Yoon et al., 2017). In our polyp dataset, biopsies and stool samples had a higher abundance of *F. prausnitzii*. However, the swabs showed a lower abundance of this taxon. Thus, we did not find a consistent pattern for the change of *F. prausnitzii* with respect to adenoma and CRC based on these studies.

5.5 Finding common OTUs among three datasets

In our analysis of two benchmarks and MBO1 polyp dataset, there were some OTUs that showed a significant change in all three datasets such as genera *Bacteroides*, *Blautia*, *Oscillospira, Ruminococcus,* and *Suterrella* at the genus level. At the species level, we observed that *Bacteroides fragilis* was significantly different between the groups in all three datasets as showed in Table 28, however, the direction of change was not the same as it showed in Table 29. We discussed the function of *Bacteroides* and *Blautia* in the previous section and more information will be provided about the rest of these ubiquitous genera here. The direction of changes of these taxa is shown in Table 29.

Table 28 The OTUs that showed a significant change in all three datasets of Benchmarks 1, 2, and MBO1 polyp dataset.
One OTU at the species level, seven at the genus level, two at the family level, and one at the order level were detected.

No.	OTU taxonomy
1	p_Firmicutes;c_Clostridia;o_Clostridiales;
2	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;
3	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;
4	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;
5	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides;s_fragilis
6	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia;
7	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira;
8	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus;
9	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Ruminococcus;
10	p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Sutterella;
11	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides;

Oscillospira*: Oscillospira* is an anaerobic bacterial genus from *Clostridial* cluster IV that has been detected in gut microbiota in several recent 16s rRNA studies and was associated with some traits like leanness (Makivuokko et al., 2010). This genus is less abundant in patients with inflammatory bowel disease compared to healthy control individuals (Walters et al., 2014). *Oscillospira* species are butyrate producers, and some of these species may digest glucuronate common in a meat-based diet and also produced by human cells. Some of the *Oscillospira* species may also digest host glycans and produce butyrate. In type II diabetes, a reduction in butyrate-producing bacteria has also been reported (Qin et al., 2012; Karlsson et al., 2013). Butyrate may also be critical in metabolic diseases (Arora and Backhed, 2016) and it has been reported to be reduced in several

inflammatory diseases (Zhu et al., 2013; Walters et al., 2014). Thus, as a butyrate producer, *Oscillospira* may be very important to human health (Gophna et al., 2017).

In Benchmark-1, *Oscillospira* was significantly lower in adenoma than healthy control individuals, but the difference between the cancer group and the healthy control group was not significant. In contrast, this genus showed a significant increase in the cancer group compared to the healthy control group in Benchmark-2. In the biopsy and stool samples collected from patients of the polyp-Y cohort, *Oscillospira* abundance was significantly reduced, however, in rectal swabs taken from the same cohort, one *Oscillospira* OTU decreased while another increased. Again, there is no consistent pattern for this genus.

Ruminococcus: *Ruminococcus* is a gram-positive bacterium that resides in the human gut and can digest resistant starches and complex carbohydrates in high fiber foods like lentils, beans, and unprocessed whole grains (Ze et al., 2012). *Ruminococcus bromii* is found to be increased in abundance in the microbiome of individuals with resistant starch diet (Walker et al., 2011; Ze et al., 2012). The slow digestion of these particular carbohydrates by *Ruminococcus* has been associated with numerous health benefits such as reversing infectious diarrhea, reducing the risk of diabetes, and preventing colon cancer (Ramakrishna et al., 2000; Niderman-Meyer et al., 2010; Robertson et al., 2005; Young et al., 2005; Le Leu et al., 2009). One species of *Ruminococcus* has been associated with increased severity of irritable bowel syndrome, but most species are vital and necessary for healthy digestive function (Malinen et al., 2010).

In the Firmicutes phylum, the genus *Ruminococcus* has a very high phylogenetic diversity (Rajilic-Stojanovic & De Vos, 2014) and there are many misclassified species in this genus. Some *Ruminococcus* species were recently reclassified as the genus *Blautia* (Liu et al., 2008).

We observed that there was a reduction of *Ruminococcus* in cancer and adenoma samples compared to the healthy control group in Benchmark-1, whereas in Benchmark-2, this genus showed an increase in the cancer group. In biopsy samples of the polyp study, there were four *Ruminococcus* OTUs that were higher in the polyp-Y group and one OTU with a lower abundance in the polyp-Y group. However, in stool and rectal swab samples, *Ruminococcus* was higher in the polyp-Y group. Therefore, the pattern of change is not the same in all of these datasets.

Sutterella: *Sutterella* species are Gram-negative, anaerobic or microaerophilic rods, are bile-resistant, and are asaccharolytic, (Wexler, 2005). In some individuals, *Sutterella* is a normal part of the microbiota. However, *Sutterella* has also been detected in intestinal biopsy and stool of patients with Crohn's disease and ulcerative colitis (Mangin et al., 2004; Gophna et al., 2006). In a comprehensive study on adenoma, *Sutterella* has been reported to be a significantly enriched taxon in the adenoma group and it was proposed as one of the four taxa that can predict adenomatous polyps (Hale et al., 2017). In contrast, another study reported a lower abundance of *Sutterella wadsworthia* in adenoma stool samples compared to the healthy control group (Brim et al., 2013). The genus *Sutterella* was higher in the cancer group compared to the adenoma group in Benchmark-1. However, in Benchmark-2, *Sutterella* was lower in the cancer state with

respect to the healthy control subjects. In polyp dataset, biopsies showed a higher abundance of *Sutterella* in the polyp-Y group while rectal swabs had a lower abundance of this genus in the polyp-Y group. Again, there is lack of consistency in the observations among different studies.

Table 29 The direction of change for common significant OTUs among three studies. It is hard to find a regular pattern of change for these taxa as different studies are not the same regards to the technical aspects and recruited population. H: Healthy; Ad: Adenoma; Ca: Cancer; PY: polyp-Y; PN: polyp-N.

	Benchmark-1			Benchmark-2	MBO1 polyp dataset		
Groups	**Ad-H**	**Ca-H**	**Ad-Ca**	**Ca-H**	**Biopsy**	**Stool**	**swabs**
Bacteroides	Ad<H Ad>H	Ca<H	Ca>Ad	Ca<H Ca>H	PY>PN	PY>PN	PY>PN
Bacteroides fragilis	N/S	N/S	Ca>Ad	Ca>H	PY<PN	N/S	N/S
Blautia	Ad>H	Ca<H	Ca<Ad	Ca<H	PY<PN	PY>PN	PY>PN
Oscillospira	Ad<H	N/S	N/S	Ca>H	PY<PN	PY<PN	PY>PN PY<PN
f_Ruminococcaceae; ***g_Ruminococcus***	Ad<H	Ca<H	Ca<Ad	Ca>H	PY<PN PY>PN	PY>PN	PY>PN
f_Lachnospiraceae; ***g_Ruminococcus;***	N/S	N/S	N/S	Ca>H	PY<PN	N/S	PY>PN
Sutterella	N/S	N/S	Ca>Ad	Ca<H	PY>PN	N/S	PY<PN
Parabacteroides	Ad<H	N/S	N/S	Ca>H	N/S	N/S	PY>PN

5.6 Strengths and weaknesses of this study

Like any research project, there are strengths and weaknesses for this study. This study was a clinical trial with human subjects that brings up several limitations to keep everything as planned and ultimately collect all type of sample and clinical information.

A few notable strengths are:

1. Large sample size: compared to other CRC/ adenoma microbiome studies, we had a superior number of participants (218 subjects in total which may have reduced the rate of type II errors (incorrectly retaining a false null hypothesis).
2. Three different sample types: biopsy, stool, and rectal swabs were collected from the subjects. We were able to show that if the sample type is different, the diversity and abundance patterns may not be the same.
3. Using three OTU clustering approaches: the most widely used OTU clustering approaches that are available in QIIME, Mothur, and USEARCH were used in this study, and the results were compared to each other. We determined that USEARCH was the preferred method.
4. Four statistical tests were used in this study to help to extract statistically significant OTUs for downstream analysis. As many of the OTUs were not significantly changed between disease and healthy states, removing them from the further analysis reduced the computational expense and improved the accuracy of classification and identification of bacteria that are associated with the disease condition.
5. More than one classification approach was used. Therefore, we determined which of these classification methods could make better performing classifiers.

Some of the limitations of this project are:

1. The clinical information about our subjects was not complete. We did not have diet information so we could not assess the effects of diet on the microbiome disease associations.
2. The design was cross-sectional, which precludes analysis of temporality and conclusions about etiology rather than mere associations.
3. More statistical tests and machine learning methods are available that could be used to find efficient OTUs in polypogenesis and predict classifiers for unknown samples.

6 CONCLUSIONS

In this dissertation, two published studies of colorectal cancer and adenoma, as well as a dataset from our polyp study were analyzed using three fundamentally different analytic pipelines: UPARSE, UPGMA, and UCLUST. These pipelines differ from each other with respect to their speed, memory usage, the number of detected OTUs, and the taxonomy of the OTUs. In addition, the number of significant OTUs detected by each pipeline was different. Among these three clustering approaches, UPARSE was faster than the other two methods, and UCLUST was the slowest. In addition, UPARSE returned the lowest number of OTUs and UCLUST the highest. Alpha and beta diversity outputs of the pipelines differed by actual diversity values, but adequately described general diversity trends. Use of OTUs identified by the UPARSE algorithm allowed the derivation of better-performing classifiers as compared to OTUs extracted using either UCLUST or UPGMA.

In conclusion, each of these analytic approaches has its advantages and disadvantages. Therefore, researchers should select an appropriate analytic pipeline depending on the nature of their datasets and the study aims. In general, UPARSE is the preferable clustering algorithm of three algorithms compared. Moreover, our study supports the current understanding that UCLUST outputs are more or less stochastic.

We hypothesized that prior feature selection would improve classification accuracy using machine learning. When only OTUs that significantly changed between binary

groups were used as inputs, the machine-learning algorithms generated classifiers that performed better than those generated using all detected OTUs without pre-selection. Therefore, we recommend adding statistical preselection steps to the 16S analytic pipelines. In this way, the number of features is reduced to the most informative OTUs. As an added benefit, an analysis of reduced dataset would require less computational power.

Three different types of samples were collected for the polyp malignancy study: the biopsy, the rectal swab, and the stool. When individuals with polyps were compared to ones with normal colonoscopy results, their microbial profiles were different in all three types of the specimens. Even for the same individual, the microbial profiles of their biopsy, stool, and rectal swabs differed. These observations confirm ones made in previous studies (Chen et al., 2012; Mira-Pascual et al., 2015).

We expected to see that the comparison of OTUs (i.e., taxa) reported in previously published studies and those identified in our own work would have many common species and they would change their abundance along with the appearance of polyps. Indeed, some bacterial taxa highlighted by our study were previously reported to be associated with adenoma/CRC. These include the Firmicutes, Bacteroidetes, and Proteobacteria phyla, *Bacteroides*, *Roseburia*, *Bifidobacterium*, *Faecalibacterium*, and *Blautia* genera, as well as *Bacteroides fragilis* and *Faecalibacterium prausnitzii* species. However, the direction of change was not collinear between in all the studies. One possible explanation for this phenomenon is the difference in species and strain composition between study populations and partial overlap between the biochemical functions of particular microbes. Another possible explanation is the presence of certain clinical confounders (e.g., the differences in

BMI, age, and medications), ethnic backgrounds, dietary habits, or last, but not least, by the differences in the sequencing methods employed, 16S variable region analyzed, and analytic pipelines which heavily influence the study outputs. Even more peculiar, the patterns of observed changes differed depending on kind of collected sample: the biopsy, the rectal swab, or stool.

Roseburia, *Bifidobacterium*, *Faecalibacterium*, and *Blautia* are commensal bacteria that produce short-chain fatty acids, particularly the butyrate. Speaking generally, these bacteria are beneficial for the body as they play a role in colonic motility, immunity maintenance and anti-inflammatory responses (Tamanai-Shacoori et al., 2017). In the studies of CRC and adenoma cohorts, there is no consensus on enrichment or depletion of these microorganisms. On one hand, the progression of a polyp toward becoming malignant may be accompanied by genuine shifts in abundance of these beneficial bacterial taxa. On the other hand, the gut microbiome may shift to increase the number of beneficial bacteria in order to maintain homeostasis in the failing mucosal barrier, thus explaining disease associated increases in abundance of these beneficial genera in the disease state reported in some of the studies. In MBO1 polyp data set, the abundancies of *Bifidobacterium*, *Faecalibacterium*, and *Blautia* are decreased in polyp biopsies as compared to normal colon mucosa, while in the rectal swabs and stool samples of the patients with polyps, their abundance (except for *Faecalibacterium*) were higher than that in patients with healthy colons. Possibly, the polyp-associated gut microbiota shifts relative abundancies toward more harmful bacteria, while beneficial bacteria are displaced into the lumen and gradually shed with the stool.

In our study, in patients with the polyps, pathogens like *Fusobacterium nucleatum* and *Bacteroides fragilis* did not show enrichment. It seems that in our cohort the polypogenesis was not associated with these two pathogens, enriched in many other CRC datasets. As the origin and etiology of CRC may be different based on the tumor location (Lee et al., 2015; Petere et al., 2016), it is possible that *Fusobacterium nucleatum* and *Bacteroides fragilis* contribute to the development of the polyps or malignant transformations of adenomatous polyps in some locations, but not others. Another confounding factor is that some of the patients in our study were so-called "polyp producers" which were monitored annually for a removal of new polyps. Due to the possible genetic component, the etiology of the polyps in this group probably differed that in the general population.

The differences in the spectrum of observed OTUs reported in the literature may be explained by many possible reasons including technical and populational variability and the sample size. Some previously published studies had an insufficient number of subjects that may lead to spurious results. For example, Marchesi et al. 2011 analyzed only six pairs of CRC and the adjacent normal tissues, and Mira-Pascual et al. 2015 had fecal and biopsy samples collected from seven patients with CRC, 11 with tubular adenomas, and ten were healthy control subjects. Clearly, there are few studies with statistically confirmed power to detect the difference in abundancies in these pathologies. It should also be noted that assigning OTUs to the same taxa does not mean those OTUs necessarily have the same function, while the differences in relative taxa abundancies do not necessarily mean that the microbial communities are functionally different. Because we chose 97% identity as

the threshold for assigning reads into OTUs, we should consider that even two reads that fall in the same category are not 100% similar and could be different strains. Even if they are 100% similar with respect to the 16S rRNA, we cannot be sure that they are functionally the same. Sometimes a single species may have two 16S genes with less than 97% similarity, or two different species may have 16S sequences with more than 99% similarity (Navas-Molina et al., 2013).

In any case, microbiome studies of human adenomas and CRC reveal useful information about the bacterial structure of the intestinal microbiome and its changes along with the progression of the disease. However, this field is very far form maturing into a clinical diagnostic and/or prognostic approach. The most crucial step is standardizing the process of microbiota assessment, starting from sample collection and storage, nucleic acid extraction, sequencing parameters, and post-sequencing analytics for processing the reads and selecting OTUs-based predictors. Thus, the continued evolution of microbiome analysis field would proceed along with further development of sequencing technology and data processing algorithms. It is also likely that much larger cohorts would be required in order to account for clinical confounders and ethnic differences in abundancies of various microbial species.

In summary, this dissertation focused on the comparative investigation of the statistical analysis and machine learning methods applicable in colorectal microbiome studies. By optimizing analytic pipeline, we were able to improve extraction of informative OTUs, thus, also improving the predicting power of resultant classifier models in sorting out the condition of the human colon. It is possible that the best classifying features are

also important in the etiology of colorectal cancer. However, before microbiome-based classifiers could be introduced to clinical practice, further validation in independent cohorts should be performed.

REFERENCES

Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj and M. F. Polz (2004). "Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons." Journal of bacteriology **186**(9): 2629-2635.

Ahlquist, D. A. and A. P. Shuber (2002). "Stool screening for colorectal cancer: evolution from occult blood to molecular markers." Clin Chim Acta **315**(1-2): 157-168.

Ahlquist, D. A., H. S. Wieand, C. G. Moertel, D. B. McGill, C. L. Loprinzi, M. J. O'Connell, J. A. Mailliard, J. B. Gerstner, K. Pandya and R. D. Ellefson (1993). "Accuracy of fecal occult blood screening for colorectal neoplasia. A prospective study using Hemoccult and HemoQuant tests." JAMA **269**(10): 1262-1267.

Ahmed, F. E., N. C. Ahmed, P. W. Vos, C. Bonnerup, J. N. Atkins, M. Casey, G. J. Nuovo, W. Naziri, J. E. Wiley and H. Mota (2013). "Diagnostic microRNA markers to screen for sporadic human colon cancer in stool: I. Proof of principle." Cancer Genomics-Proteomics **10**(3): 93-113.

Ahn, J., R. Sinha, Z. Pei, C. Dominianni, J. Wu, J. Shi, J. J. Goedert, R. B. Hayes and L. Yang (2013). "Human gut microbiome and risk for colorectal cancer." J Natl Cancer Inst **105**(24): 1907-1911.

Allison, J. E., L. C. Sakoda, T. R. Levin, J. P. Tucker, I. S. Tekawa, T. Cuff, M. P. Pauly, L. Shlager, A. M. Palitz and W. K. Zhao (2007). "Screening for colorectal neoplasms with new fecal occult blood tests: update on performance characteristics." Journal of the National Cancer Institute **99**(19): 1462-1470.

Allison, J. E., I. S. Tekawa, L. J. Ransom and A. L. Adrain (1996). "A comparison of fecal occult-blood tests for colorectal-cancer screening." N Engl J Med **334**(3): 155-159.

Amitay, E. L., S. Werner, M. Vital, D. H. Pieper, D. Hofler, I. J. Gierse, J. Butt, Y. Balavarca, K. Cuk and H. Brenner (2017). "Fusobacterium and colorectal cancer: Causal factor or passenger? Results from a large colorectal cancer screening study." Carcinogenesis.

Arora, T. and F. Backhed (2016). "The gut microbiota and metabolic disease: current understanding and future perspectives." J Intern Med **280**(4): 339-349.

Arthur, J. C., E. Perez-Chanona, M. Mühlbauer, S. Tomkovich, J. M. Uronis, T.-J. Fan, B. J. Campbell, T. Abujamel, B. Dogan and A. B. Rogers (2012). "Intestinal inflammation targets cancer-inducing activity of the microbiota." science **338**(6103): 120-123.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J. M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dore, H. I. T. C. Meta, M. Antolin, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Merieux, R. Melo Minardi, C. M'Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich and P. Bork (2011). "Enterotypes of the human gut microbiome." Nature **473**(7346): 174-180.

Bachy, C., J. R. Dolan, P. López-García, P. Deschamps and D. Moreira (2013). "Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study." The ISME journal **7**(2): 244-255.

Baxter, N. N., M. A. Goldwasser, L. F. Paszat, R. Saskin, D. R. Urbach and L. Rabeneck (2009). "Association of colonoscopy and death from colorectal cancer." Annals of internal medicine **150**(1): 1-8.

Baxter, N. T., J. P. Zackular, G. Y. Chen and P. D. Schloss (2014). "Structure of the gut microbiome following colonization with human feces determines colonic tumor burden." Microbiome **2**(1): 20.

Beck, D. and J. A. Foster (2014). "Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics." PLoS One **9**(2): e87830.

Beck, D. and J. A. Foster (2015). "Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis." BioData Min **8**: 23.

Beeckman, D. S. and D. C. Vanrompay (2010). "Bacterial secretion systems with an emphasis on the chlamydial Type III secretion system." Curr Issues Mol Biol **12**(1): 17-41.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler (2005). "GenBank." Nucleic acids research **33**(suppl 1): D34-D38.

Bianconi, E., A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani and F. Piva (2013). "An estimation of the number of cells in the human body." Annals of human biology **40**(6): 463-471.

Bosch, L. J. W., B. Carvalho, R. J. A. Fijneman, C. R. Jimenez, H. M. Pinedo, M. van Engeland and G. A. Meijer (2011). "Molecular tests for colorectal cancer screening." Clinical colorectal cancer **10**(1): 8-23.

Bottacini, F., M. Ventura, D. van Sinderen and M. O'Connell Motherway (2014). "Diversity, ecology and intestinal function of bifidobacteria." Microb Cell Fact **13 Suppl 1**: S4.

Brady, A. and S. L. Salzberg (2009). "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models." Nature methods **6**(9): 673-676.

Bragg, L., G. Stone, M. Imelfort, P. Hugenholtz and G. W. Tyson (2012). "Fast, accurate error-correction of amplicon pyrosequences using Acacia." Nature methods **9**(5): 425-426.

Bray, J. R. and J. T. Curtis (1957). "An ordination of the upland forest communities of southern Wisconsin." Ecological monographs **27**(4): 325-349.

Brenner, H., J. Chang-Claude, C. M. Seiler, A. Rickert and M. Hoffmeister (2011). "Protection from colorectal cancer after colonoscopy: a population-based, case–control study." Annals of internal medicine **154**(1): 22-30.

Brenner, H. and S. Tao (2013). "Superior diagnostic performance of faecal immunochemical tests for haemoglobin in a head-to-head comparison with guaiac based faecal occult blood test among 2235 participants of screening colonoscopy." Eur J Cancer **49**(14): 3049-3054.

Brim, H., S. Yooseph, E. G. Zoetendal, E. Lee, M. Torralbo, A. O. Laiyemo, B. Shokrani, K. Nelson and H. Ashktorab (2013). "Microbiome analysis of stool samples from African Americans with colon polyps." PloS one **8**(12): e81352.

Brown, N. F. and B. B. Finlay (2011). "Potential origins and horizontal transfer of type III secretion systems and effectors." Mob Genet Elements **1**(2): 118-121.

Cai, Y. and Y. Sun (2011). "ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time." Nucleic Acids Res **39**(14): e95.

Candela, M., S. Turroni, E. Biagi, F. Carbonero, S. Rampelli, C. Fiorentini and P. Brigidi (2014). "Inflammation and colorectal cancer, when microbiota-host mutualism breaks." World journal of gastroenterology: WJG **20**(4): 908.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich and J. I. Gordon (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature methods **7**(5): 335-336.

Carmona, F. J., D. Azuara, A. Berenguer-Llergo, A. F. Fernandez, S. Biondo, J. de Oca, F. Rodriguez-Moranta, R. Salazar, A. Villanueva, M. F. Fraga, J. Guardiola, G. Capella, M. Esteller and V. Moreno (2013). "DNA methylation biomarkers for noninvasive diagnosis of colorectal cancer." Cancer Prev Res (Phila) **6**(7): 656-665.

Carvalho, F. A., O. Koren, J. K. Goodrich, M. E. V. Johansson, I. Nalbantoglu, J. D. Aitken, Y. Su, B. Chassaing, W. A. Walters and A. González (2012). "Transient inability to manage proteobacteria promotes chronic gut inflammation in TLR5-deficient mice." Cell host & microbe **12**(2): 139-152.

Castellarin, M., R. L. Warren, J. D. Freeman, L. Dreolini, M. Krzywinski, J. Strauss, R. Barnes, P. Watson, E. Allen-Vercoe and R. A. Moore (2012). "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma." Genome research **22**(2): 299-306.

Chang, P. V., L. Hao, S. Offermanns and R. Medzhitov (2014). "The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition." Proc Natl Acad Sci U S A **111**(6): 2247-2252.

Chen, W., F. Liu, Z. Ling, X. Tong and C. Xiang (2012). "Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer." PLoS One **7**(6): e39743.

Chen, W., C. K. Zhang, Y. Cheng, S. Zhang and H. Zhao (2013). "A comparison of methods for clustering 16S rRNA sequences into OTUs." PloS one **8**(8): e70837.

Cheng, T. I., J. M. Wong, C. F. Hong, S. H. Cheng, T. J. Cheng, M. J. Shieh, Y. M. Lin, C. Y. Tso and A. T. Huang (2002). "Colorectal cancer screening in asymptomaic adults: comparison of colonoscopy, sigmoidoscopy and fecal occult blood tests." J Formos Med Assoc **101**(10): 685-690.

Chiang, T. H., Y. C. Lee, C. H. Tu, H. M. Chiu and M. S. Wu (2011). "Performance of the immunochemical fecal occult blood test in predicting lesions in the lower gastrointestinal tract." CMAJ **183**(13): 1474-1481.

Chiu, H. M., Y. C. Lee, C. H. Tu, C. C. Chen, P. H. Tseng, J. T. Liang, C. T. Shun, J. T. Lin and M. S. Wu (2013). "Association between early stage colon neoplasms and false-negative results from the fecal immunochemical test." Clin Gastroenterol Hepatol **11**(7): 832-838 e831-832.

Cho, E., S. A. Smith-Warner, D. Spiegelman, W. L. Beeson, P. A. van den Brandt, G. A. Colditz, A. R. Folsom, G. E. Fraser, J. L. Freudenheim and E. Giovannucci (2004). "Dairy foods, calcium, and colorectal cancer: a pooled analysis of 10 cohort studies." Journal of the National Cancer Institute **96**(13): 1015-1022.

Clarridge, J. E. (2004). "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases." Clinical microbiology reviews **17**(4): 840-862.

Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity and J. M. Tiedje (2007). "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data." Nucleic acids research **35**(suppl 1): D169-D172.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh and G. M. Garrity (2009). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." Nucleic acids research **37**(suppl 1): D141-D145.

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske and J. M. Tiedje (2014). "Ribosomal Database Project: data and tools for high throughput rRNA analysis." Nucleic Acids Res **42**(Database issue): D633-642.

Collins, S. M. (2014). "A role for the gut microbiota in IBS." Nature reviews Gastroenterology & hepatology **11**(8): 497-505.

Costello, E. K., J. I. Gordon, S. M. Secor and R. Knight (2010). "Postprandial remodeling of the gut microbiota in Burmese pythons." ISME J **4**(11): 1375-1385.

Cushing, K., D. M. Alvarado and M. A. Ciorba (2015). "Butyrate and Mucosal Inflammation: New Scientific Evidence Supports Clinical Observation." Clin Transl Gastroenterol **6**: e108.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma and M. A. Fischbach (2014). "Diet rapidly and reproducibly alters the human gut microbiome." Nature **505**(7484): 559-563.

De Angelis, M., M. Piccolo, L. Vannini, S. Siragusa, A. De Giacomo, D. I. Serrazzanetti, F. Cristofori, M. E. Guerzoni, M. Gobbetti and R. Francavilla (2013). "Fecal microbiota and metabolome of children with autism and pervasive developmental disorder not otherwise specified." PLoS One **8**(10): e76993.

De Filippo, C., D. Cavalieri, M. Di Paola, M. Ramazzotti, J. B. Poullet, S. Massart, S. Collini, G. Pieraccini and P. Lionetti (2010). "Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa." Proceedings of the National Academy of Sciences **107**(33): 14691-14696.

de Wijkerslooth, T. R., E. M. Stoop, P. M. Bossuyt, G. A. Meijer, M. van Ballegooijen, A. H. van Roon, I. Stegeman, R. A. Kraaijenhagen, P. Fockens, M. E. van Leerdam, E. Dekker and E. J. Kuipers (2012). "Immunochemical fecal occult blood testing is equally sensitive for proximal and distal advanced neoplasia." Am J Gastroenterol **107**(10): 1570-1578.

Demšar, J., T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak and A. Starič (2013). "Orange: data mining toolbox in Python." The Journal of Machine Learning Research **14**(1): 2349-2353.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Applied and environmental microbiology **72**(7): 5069-5072.

Dethlefsen, L. and D. A. Relman (2011). "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation." Proceedings of the National Academy of Sciences **108**(Supplement 1): 4554-4561.

Dominguez-Bello, M. G., E. K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer and R. Knight (2010). "Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns." Proceedings of the National Academy of Sciences **107**(26): 11971-11975.

Dubé, C., A. Rostom, G. Lewin, A. Tsertsvadze, N. Barrowman, C. Code, M. Sampson and D. Moher (2007). "The use of aspirin for primary prevention of colorectal cancer: a systematic review prepared for the US Preventive Services Task Force." Annals of internal medicine **146**(5): 365-375.

Dufrêne, M. and P. Legendre (1997). "Species assemblages and indicator species: the need for a flexible asymmetrical approach." Ecological monographs **67**(3): 345-366.

Dulal, S. and T. O. Keku (2014). "Gut microbiome and colorectal adenomas." Cancer J **20**(3): 225-231.

Dupuy, A. and R. M. Simon (2007). "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting." J Natl Cancer Inst **99**(2): 147-157.

Duranti, S., C. Milani, G. A. Lugli, L. Mancabelli, F. Turroni, C. Ferrario, M. Mangifesta, A. Viappiani, B. Sanchez, A. Margolles, D. van Sinderen and M. Ventura (2016). "Evaluation of genetic diversity among strains of the human gut commensal Bifidobacterium adolescentis." Sci Rep **6**: 23971.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Edgar, R. C. and H. Flyvbjerg (2015). "Error filtering, pair assembly and error correction for next-generation sequencing reads." Bioinformatics **31**(21): 3476-3482.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince and R. Knight (2011). "UCHIME improves sensitivity and speed of chimera detection." Bioinformatics **27**(16): 2194-2200.

Edwards, B. K., E. Ward, B. A. Kohler, C. Eheman, A. G. Zauber, R. N. Anderson, A. Jemal, M. J. Schymura, I. Lansdorp-Vogelaar and L. C. Seeff (2010). "Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates." Cancer **116**(3): 544-573.

Eklof, V., A. Lofgren-Burstrom, C. Zingmark, S. Edin, P. Larsson, P. Karling, O. Alexeyev, J. Rutegard, M. L. Wikberg and R. Palmqvist (2017). "Cancer associated faecal microbial markers in colorectal cancer detection." Int J Cancer.

Faivre, J., V. Dancourt, C. Lejeune, M. A. Tazi, J. Lamour, D. Gerard, F. Dassonville and C. Bonithon-Kopp (2004). "Reduction in colorectal cancer mortality by fecal occult blood screening in a French controlled study." Gastroenterology **126**(7): 1674-1680.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). "From data mining to knowledge discovery in databases." AI magazine **17**(3): 37.

Fearon, E. R. (2011). "Molecular genetics of colorectal cancer." Annual Review of Pathology: Mechanisms of Disease **6**: 479-507.

Fearon, E. R. and B. Vogelstein (1990). "A genetic model for colorectal tumorigenesis." Cell **61**(5): 759-767.

Ferlay, J., I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman and F. Bray (2015). "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." International journal of cancer **136**(5).

Findley, K., J. Oh, J. Yang, S. Conlan, C. Deming, J. A. Meyer, D. Schoenfeld, E. Nomicos, M. Park and N. I. H. I. S. C. C. Sequencing (2013). "Topographic diversity of fungal and bacterial communities in human skin." Nature **498**(7454): 367-370.

Flanagan, L., J. Schmid, M. Ebert, P. Soucek, T. Kunicka, V. Liska, J. Bruha, P. Neary, N. Dezeeuw and M. Tommasino (2014). "Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome." European journal of clinical microbiology & infectious diseases **33**(8): 1381-1390.

Flint, H. J., K. P. Scott, P. Louis and S. H. Duncan (2012). "The role of the gut microbiota in nutrition and health." Nature Reviews Gastroenterology and Hepatology **9**(10): 577-589.

Flynn, J. M., E. A. Brown, F. J. Chain, H. J. MacIsaac and M. E. Cristescu (2015). "Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods." Ecology and evolution **5**(11): 2252-2266.

Fraher, M. H., P. W. O'Toole and E. M. M. Quigley (2012). "Techniques used to characterize the gut microbiota: a guide for the clinician." Nature Reviews Gastroenterology and Hepatology **9**(6): 312-322.

Frank, D. N., A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz and N. R. Pace (2007). "Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases." Proc Natl Acad Sci U S A **104**(34): 13780-13785.

Furusawa, Y., Y. Obata, S. Fukuda, T. A. Endo, G. Nakato, D. Takahashi, Y. Nakanishi, C. Uetake, K. Kato, T. Kato, M. Takahashi, N. N. Fukuda, S. Murakami, E. Miyauchi, S. Hino, K. Atarashi, S. Onawa, Y. Fujimura, T. Lockett, J. M. Clarke, D. L. Topping, M. Tomita, S. Hori, O. Ohara, T. Morita, H. Koseki, J. Kikuchi, K. Honda, K. Hase and H. Ohno (2013). "Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells." Nature **504**(7480): 446-450.

Gao, R., C. Kong, L. Huang, H. Li, X. Qu, Z. Liu, P. Lan, J. Wang and H. Qin (2017). "Mucosa-associated microbiota signature in colorectal cancer." Eur J Clin Microbiol Infect Dis **36**(11): 2073-2083.

Geng, J., H. Fan, X. Tang, H. Zhai and Z. Zhang (2013). "Diversified pattern of the human colorectal cancer microbiome." Gut Pathog **5**(2).

Ghodsi, M., B. Liu and M. Pop (2011). "DNACLUST: accurate and efficient clustering of phylogenetic marker genes." BMC Bioinformatics **12**: 271.

Gill, S. R., M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett and K. E. Nelson (2006). "Metagenomic analysis of the human distal gut microbiome." science **312**(5778): 1355-1359.

Gillevet, P. M. (2008). Multitag sequencing and ecogenomics analysis, Google Patents.

Giovannucci, E. (2002). "Modifiable risk factors for colon cancer." Gastroenterology Clinics of North America **31**(4): 925-943.

Glass, E. M., J. Wilkening, A. Wilke, D. Antonopoulos and F. Meyer (2010). "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes." Cold Spring Harbor Protocols (2010): pdb. prot5368.

Goedert, J. J., Y. Gong, X. Hua, H. Zhong, Y. He, P. Peng, G. Yu, W. Wang, J. Ravel, J. Shi and Y. Zheng (2015). "Fecal Microbiota Characteristics of Patients with Colorectal Adenoma Detected by Screening: A Population-based Study." EBioMedicine **2**(6): 597-603.

Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters, J. G. Caporaso, R. Knight and R. E. Ley (2014). "Conducting a microbiome study." Cell **158**(2): 250-262.

Goodrich, J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark and R. E. Ley (2014). "Human genetics shape the gut microbiome." Cell **159**(4): 789-799.

Gophna, U., T. Konikoff and H. B. Nielsen (2017). "Oscillospira and related bacteria - From metagenomic species to metabolic features." Environ Microbiol **19**(3): 835-841.

Gophna, U., K. Sommerfeld, S. Gophna, W. F. Doolittle and S. J. Veldhuyzen van Zanten (2006). "Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis." J Clin Microbiol **44**(11): 4136-4141.

Gray, M. W., G. Burger and B. F. Lang (1999). "Mitochondrial evolution." Science **283**(5407): 1476-1481.

Grodstein, F., P. A. Newcomb and M. J. Stampfer (1999). "Postmenopausal hormone therapy and the risk of colorectal cancer: a review and meta-analysis." The American journal of medicine **106**(5): 574-582.

Guarner, F. and J. R. Malagelada (2003). "Gut flora in health and disease." Lancet **361**(9356): 512-519.

Haas, B. J., D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, D. Ciulla, D. Tabbaa, S. K. Highlander and E. Sodergren (2011). "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons." Genome research **21**(3): 494-504.

Hale, V. L., J. Chen, S. Johnson, S. C. Harrington, T. C. Yab, T. C. Smyrk, H. Nelson, L. A. Boardman, B. R. Druliner, T. R. Levin, D. K. Rex, D. J. Ahnen, P. Lance, D. A. Ahlquist and N. Chia (2017). "Shifts in the Fecal Microbiota Associated with Adenomatous Polyps." Cancer Epidemiol Biomarkers Prev **26**(1): 85-94.

Hardcastle, J. D., J. O. Chamberlain, M. H. Robinson, S. M. Moss, S. S. Amar, T. W. Balfour, P. D. James and C. M. Mangham (1996). "Randomised controlled trial of faecal-occult-blood screening for colorectal cancer." Lancet **348**(9040): 1472-1477.

Heijtz, R. D., S. Wang, F. Anuar, Y. Qian, B. Björkholm, A. Samuelsson, M. L. Hibberd, H. Forssberg and S. Pettersson (2011). "Normal gut microbiota modulates brain development and behavior." Proceedings of the National Academy of Sciences **108**(7): 3047-3052.

Hill, M. O. (1973). "Diversity and evenness: a unifying notation and its consequences." Ecology **54**(2): 427-432.

Hold, G. L., M. Smith, C. Grange, E. R. Watt, E. M. El-Omar and I. Mukhopadhya (2014). "Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years?" World journal of gastroenterology: WJG **20**(5): 1192.

Hooper, L. V., D. R. Littman and A. J. Macpherson (2012). "Interactions between the microbiota and the immune system." Science **336**(6086): 1268-1273.

Imperiale, T. F., D. F. Ransohoff, S. H. Itzkowitz, T. R. Levin, P. Lavin, G. P. Lidgard, D. A. Ahlquist and B. M. Berger (2014). "Multitarget stool DNA testing for colorectal-cancer screening." New England Journal of Medicine **370**(14): 1287-1297.

Imperiale, T. F., D. F. Ransohoff, S. H. Itzkowitz, B. A. Turnbull, M. E. Ross and G. Colorectal Cancer Study (2004). "Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population." N Engl J Med **351**(26): 2704-2714.

Issa, I. A. and M. Noureddine (2017). "Colorectal cancer screening: An updated review of the available options." World J Gastroenterol **23**(28): 5086-5096.

Itoh, M., K. Takahashi, H. Nishida, K. Sakagami and T. Okubo (1996). "Estimation of the optimal cut off point in a new immunological faecal occult blood test in a corporate colorectal cancer screening programme." J Med Screen **3**(2): 66-71.

Jandhyala, S. M., R. Talukdar, C. Subramanyam, H. Vuyyuru, M. Sasikala and D. Nageshwar Reddy (2015). "Role of the normal gut microbiota." World J Gastroenterol **21**(29): 8787-8803.

Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu and M. J. Thun (2009). "Cancer statistics, 2009." CA: a cancer journal for clinicians **59**(4): 225-249.

Joly, F., C. Mayeur, A. Bruneau, M. L. Noordine, T. Meylheuc, P. Langella, B. Messing, P. H. Duee, C. Cherbuy and M. Thomas (2010). "Drastic changes in fecal and mucosa-associated microbiota in adult patients with short bowel syndrome." Biochimie **92**(7): 753-761.

Jonasson, J., M. Olofsson and H. J. Monstein (2002). "Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments." Apmis **110**(3): 263-272.

Jorgensen, B. and J. Knudtson (2014). "Stop cancer colon. Colorectal cancer screening--updated guidelines." South Dakota medicine: the journal of the South Dakota State Medical Association: 82-87.

Ju, F. and T. Zhang (2015). "16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions." Applied microbiology and biotechnology **99**(10): 4119-4129.

Kahi, C. J., T. F. Imperiale, B. E. Juliar and D. K. Rex (2009). "Effect of screening colonoscopy on colorectal cancer incidence and mortality." Clinical gastroenterology and hepatology **7**(7): 770-775.

Kang, D.-W., J. G. Park, Z. E. Ilhan, G. Wallstrom, J. LaBaer, J. B. Adams and R. Krajmalnik-Brown (2013). "Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children." PLoS One **8**(7): e68322.

Kang, D. W., J. G. Park, Z. E. Ilhan, G. Wallstrom, J. Labaer, J. B. Adams and R. Krajmalnik-Brown (2013). "Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children." PLoS One **8**(7): e68322.

Karlsson, F. H., V. Tremaroli, I. Nookaew, G. Bergstrom, C. J. Behre, B. Fagerberg, J. Nielsen and F. Backhed (2013). "Gut metagenome in European women with normal, impaired and diabetic glucose control." Nature **498**(7452): 99-103.

Katoh, K., K. Misawa, K. i. Kuma and T. Miyata (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic acids research **30**(14): 3059-3066.

Keku, T. O., S. Dulal, A. Deveaux, B. Jovov and X. Han (2015). "The gastrointestinal microbiota and colorectal cancer." American Journal of Physiology-Gastrointestinal and Liver Physiology **308**(5): G351-G363.

Kindt, R. and R. Coe (2005). Tree Diversity Analysis: A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies, World Agrofirestry Centre.

Kinross, J., R. Mirnezami, J. Alexander, R. Brown, A. Scott, D. Galea, K. Veselkov, R. Goldin, A. Darzi, J. Nicholson and J. R. Marchesi (2017). "A prospective analysis of mucosal microbiome-metabonome interactions in colorectal cancer using a combined MAS 1HNMR and metataxonomic strategy." Sci Rep **7**(1): 8979.

Knights, D., E. K. Costello and R. Knight (2011). "Supervised classification of human microbiota." FEMS microbiology reviews **35**(2): 343-359.

Koenig, J. E., A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent and R. E. Ley (2011). "Succession of microbial consortia in the developing infant gut microbiome." Proceedings of the National Academy of Sciences **108**(Supplement 1): 4578-4585.

Koren, O., J. K. Goodrich, T. C. Cullender, A. Spor, K. Laitinen, H. K. Bäckhed, A. Gonzalez, J. J. Werner, L. T. Angenent and R. Knight (2012). "Host remodeling of the gut microbiome and metabolic changes during pregnancy." Cell **150**(3): 470-480.

Kostic, A. D., E. Chun, L. Robertson, J. N. Glickman, C. A. Gallini, M. Michaud, T. E. Clancy, D. C. Chung, P. Lochhead and G. L. Hold (2013). "Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment." Cell host & microbe **14**(2): 207-215.

Kostic, A. D., D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass and J. Tabernero (2012). "Genomic analysis identifies association of Fusobacterium with colorectal carcinoma." Genome research **22**(2): 292-298.

Kostic, A. D., R. J. Xavier and D. Gevers (2014). "The microbiome in inflammatory bowel disease: current status and the future ahead." Gastroenterology **146**(6): 1489-1499.

Kronborg, O., C. Fenger, J. Olsen, O. D. Jorgensen and O. Sondergaard (1996). "Randomised study of screening for colorectal cancer with faecal-occult-blood test." Lancet **348**(9040): 1467-1471.

Kuipers, E. J. (2014). "Colorectal cancer: Screening [mdash] one small step for mankind, one giant leap for man." Nature Reviews Clinical Oncology **11**(1): 5-6.

Kuipers, E. J. and A. De Jong (1990). "Gastrointestinal infection and Streptococcus bovis bacteraemia." Nederlands tijdschrift voor geneeskunde **134**(28): 1337-1339.

Kuipers, E. J., T. Rösch and M. Bretthauer (2013). "Colorectal cancer screening—optimizing current strategies and new directions." Nature Reviews Clinical Oncology **10**(3): 130-142.

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin and N. R. Pace (1985). "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses." Proceedings of the National Academy of Sciences **82**(20): 6955-6959.

Langille, M. G. I., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber and R. Knight (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." Nature biotechnology **31**(9): 814-821.

Larsson, S. C., N. Orsini and A. Wolk (2005). "Diabetes mellitus and risk of colorectal cancer: a meta-analysis." Journal of the National Cancer Institute **97**(22): 1679-1687.

Larsson, S. C. and A. Wolk (2006). "Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies." International journal of cancer **119**(11): 2657-2664.

Launoy, G. D., H. J. Bertrand, C. Berchi, V. Y. Talbourdet, A. V. Guizard, V. M. Bouvier and E. R. Caces (2005). "Evaluation of an immunochemical fecal occult blood test with automated reading in screening for colorectal cancer in a general average-risk population." Int J Cancer **115**(3): 493-496.

Le Chatelier, E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto and S. Kennedy (2013). "Richness of human gut microbiome correlates with metabolic markers." Nature **500**(7464): 541-546.

Le Leu, R. K., Y. Hu, I. L. Brown and G. P. Young (2009). "Effect of high amylose maize starches on colonic fermentation and apoptotic response to DNA-damage in the colon of rats." Nutr Metab (Lond) **6**: 11.

Lee, G., G. Malietzis, A. Askari, D. Bernardo, H. Al-Hassi and S. Clark (2015). "Is right-sided colon cancer different to left-sided colorectal cancer?–a systematic review." European Journal of Surgical Oncology (EJSO) **41**(3): 300-308.

Lee, J. K., E. G. Liles, S. Bent, T. R. Levin and D. A. Corley (2014). "Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis." Ann Intern Med **160**(3): 171.

Legendre, P. and L. F. Legendre (2012). Numerical ecology, Elsevier.

Levi, Z., S. Birkenfeld, A. Vilkin, M. Bar-Chana, I. Lifshitz, M. Chared, E. Maoz and Y. Niv (2011). "A higher detection rate for colorectal cancer and advanced adenomatous polyp for screening with immunochemical fecal occult blood test than guaiac fecal occult blood test, despite lower compliance rate. A prospective, controlled, feasibility study." Int J Cancer **128**(10): 2415-2424.

Levi, Z., P. Rozen, R. Hazazi, A. Vilkin, A. Waked, E. Maoz, S. Birkenfeld, M. Leshno and Y. Niv (2007). "A quantitative immunochemical fecal occult blood test for colorectal neoplasia." Ann Intern Med **146**(4): 244-255.

Levin, B., D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick and T. R. Levin (2008). "Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology*†." CA: a cancer journal for clinicians **58**(3): 130-160.

Ley, R. E., P. J. Turnbaugh, S. Klein and J. I. Gordon (2006). "Microbial ecology: human gut microbes associated with obesity." Nature **444**(7122): 1022-1023.

Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Liao, Z. L., B. H. Zeng, W. Wang, G. H. Li, F. Wu, L. Wang, Q. P. Zhong, H. Wei and X. Fang (2016). "Impact of the Consumption of Tea Polyphenols on Early Atherosclerotic Lesion Formation and Intestinal Bifidobacteria in High-Fat-Fed ApoE-/- Mice." Front Nutr **3**: 42.

Lieberman, D. A. (2009). "Clinical practice. Screening for colorectal cancer." N Engl J Med **361**(12): 1179-1187.

Lieberman, D. A., W. V. Harford, D. J. Ahnen, D. Provenzale, S. J. Sontag, T. G. Schnell, G. Chejfec, D. R. Campbell, T. E. Durbin and J. H. Bond (2001). "One-time screening for

colorectal cancer with combined fecal occult-blood testing and examination of the distal colon." New England Journal of Medicine **345**(8): 555-560.

Liu, C., S. M. Finegold, Y. Song and P. A. Lawson (2008). "Reclassification of Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus and Ruminococcus schinkii as Blautia coccoides gen. nov., comb. nov., Blautia hansenii comb. nov., Blautia hydrogenotrophica comb. nov., Blautia luti comb. nov., Blautia producta comb. nov., Blautia schinkii comb. nov. and description of Blautia wexlerae sp. nov., isolated from human faeces." International journal of systematic and evolutionary microbiology **58**(8): 1896-1902.

Louis, P. and H. J. Flint (2009). "Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine." FEMS Microbiol Lett **294**(1): 1-8.

Louis, P., G. L. Hold and H. J. Flint (2014). "The gut microbiota, bacterial metabolites and colorectal cancer." Nature Reviews Microbiology **12**(10): 661-672.

Louis, P., P. Young, G. Holtrop and H. J. Flint (2010). "Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA: acetate CoA-transferase gene." Environmental microbiology **12**(2): 304-314.

Lozupone, C. and R. Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities." Applied and environmental microbiology **71**(12): 8228-8235.

Lozupone, C., M. E. Lladser, D. Knights, J. Stombaugh and R. Knight (2011). "UniFrac: an effective distance metric for microbial community comparison." The ISME journal **5**(2): 169.

Lozupone, C. A., M. Hamady, S. T. Kelley and R. Knight (2007). "Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities." Applied and environmental microbiology **73**(5): 1576-1585.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, A. Buchner, T. Lai, S. Steppi, G. Jobb and W. Förster (2004). "ARB: a software environment for sequence data." Nucleic acids research **32**(4): 1363-1371.

Ma, Y., P. Zhang, F. Wang, J. Yang, Z. Liu and H. Qin (2011). "Association between vitamin D and risk of colorectal cancer: a systematic review of prospective studies." Journal of Clinical Oncology **29**(28): 3775-3782.

Makivuokko, H., K. Tiihonen, S. Tynkkynen, L. Paulin and N. Rautonen (2010). "The effect of age and non-steroidal anti-inflammatory drugs on human intestinal microbiota composition." Br J Nutr **103**(2): 227-234.

Malcomson, F. C., N. D. Willis and J. C. Mathers (2015). "Is resistant starch protective against colorectal cancer via modulation of the WNT signalling pathway?" Proc Nutr Soc **74**(3): 282-291.

Malinen, E., L. Krogius-Kurikka, A. Lyra, J. Nikkila, A. Jaaskelainen, T. Rinttila, T. Vilpponen-Salmela, A. J. von Wright and A. Palva (2010). "Association of symptoms with gastrointestinal microbiota in irritable bowel syndrome." World J Gastroenterol **16**(36): 4532-4540.

Mandel, J. S., J. H. Bond, T. R. Church, D. C. Snover, G. M. Bradley, L. M. Schuman and F. Ederer (1993). "Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study." N Engl J Med **328**(19): 1365-1371.

Mandel, J. S., T. R. Church, J. H. Bond, F. Ederer, M. S. Geisser, S. J. Mongin, D. C. Snover and L. M. Schuman (2000). "The effect of fecal occult-blood screening on the incidence of colorectal cancer." New England Journal of Medicine **343**(22): 1603-1607.

Mangin, I., R. Bonnet, P. Seksik, L. Rigottier-Gois, M. Sutren, Y. Bouhnik, C. Neut, M. D. Collins, J. F. Colombel, P. Marteau and J. Dore (2004). "Molecular inventory of faecal microflora in patients with Crohn's disease." FEMS Microbiol Ecol **50**(1): 25-36.

Manichanh, C., N. Borruel, F. Casellas and F. Guarner (2012). "The gut microbiota in IBD." Nature Reviews Gastroenterology and Hepatology **9**(10): 599-608.

Marchesi, J. R., B. E. Dutilh, N. Hall, W. H. Peters, R. Roelofs, A. Boleij and H. Tjalsma (2011). "Towards the human colorectal cancer microbiome." PloS one **6**(5): e20447.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu. Rev. Genomics Hum. Genet. **9**: 387-402.

Markowitz, V. M., N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I. M. A. Chen, Y. Grechkin, I. Dubchak and I. Anderson (2008). "IMG/M: a data management and analysis system for metagenomes." Nucleic acids research **36**(suppl 1): D534-D538.

Martinez-Medina, M., X. Aldeguer, F. Gonzalez-Huix, D. Acero and L. J. Garcia-Gil (2006). "Abnormal microbiota composition in the ileocolonic mucosa of Crohn's disease patients as revealed by polymerase chain reaction-denaturing gradient gel electrophoresis." Inflamm Bowel Dis **12**(12): 1136-1145.

McCune, B., J. B. Grace and D. L. Urban (2002). Analysis of ecological communities, MjM software design Gleneden Beach, OR.

McLean, M. H., G. I. Murray, K. N. Stewart, G. Norrie, C. Mayer, G. L. Hold, J. Thomson, N. Fyfe, M. Hope and N. A. G. Mowat (2011). "The inflammatory microenvironment in colorectal neoplasia." PLoS One **6**(1): e15366.

Mira-Pascual, L., R. Cabrera-Rubio, S. Ocon, P. Costales, A. Parra, A. Suarez, F. Moris, L. Rodrigo, A. Mira and M. C. Collado (2015). "Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers." J Gastroenterol **50**(2): 167-179.

Mohammed, M. H., T. S. Ghosh, R. M. Reddy, C. V. S. K. Reddy, N. K. Singh and S. S. Mande (2011). "INDUS-a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences." BMC genomics **12**(Suppl 3): S4.

Morgan, X. C. and C. Huttenhower (2014). "Meta'omic analytic techniques for studying the intestinal microbiome." Gastroenterology **146**(6): 1437-1448. e1431.

Morikawa, T., J. Kato, Y. Yamaji, R. Wada, T. Mitsushima and Y. Shiratori (2005). "A comparison of the immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population." Gastroenterology **129**(2): 422-428.

Nakama, H., N. Kamijo, A. S. Abdul Fattah and B. Zhang (1996). "Validity of immunological faecal occult blood screening for colorectal cancer: a follow up study." J Med Screen **3**(2): 63-65.

Nakama, H., M. Yamamoto, N. Kamijo, T. Li, N. Wei, A. S. Fattah and B. Zhang (1999). "Colonoscopic evaluation of immunochemical fecal occult blood test for detection of colorectal neoplasia." Hepatogastroenterology **46**(25): 228-231.

Nakano, Y., T. Takeshita, N. Kamio, S. Shiota, Y. Shibata, N. Suzuki, M. Yoneda, T. Hirofuji and Y. Yamashita (2014). "Supervised machine learning-based classification of oral malodor based on the microbiota in saliva samples." Artificial intelligence in medicine **60**(2): 97-101.

Nakazato, M., H.-o. Yamano, H.-o. Matsushita, K. Sato, K. Fujita, Y. Yamanaka and Y. Imai (2006). "Immunologic fecal occult blood test for colorectal cancer screening." Japan Medical Association Journal **49**(5/6): 203.

Navas-Molina, J., A., J. Peralta-Sánchez, M. , A. González, P. McMurdie, J., Y. Vázquez-Baeza, Z. Xu, L. Ursell, K., C. Lauber, H. Zhou, S. Song, Jin, J. Huntley, G. Ackermann, L., D. Berg-Lyons, S. Holmes, J. G. Caporaso and R. Knight (2013). Advancing Our Understanding of the Human Microbiome Using QIIME. Methods in Enzymology. J. N. Abelson and M. I. Simon. San Diego, CA, USA, Elsevier**:** 372-439.

Neut, C., P. Bulois, P. Desreumaux, J.-M. Membreé, E. Lederman, L. Gambiez, A. Cortot, P. Quandalle, H. Van Kruiningen and J.-F. Colombel (2002). "Changes in the bacterial flora of the neoterminal ileum after ileocolonic resection for Crohn's disease." The American journal of gastroenterology **97**(4): 939-946.

Newcomb, P. A., R. G. Norfleet, B. E. Storer, T. S. Surawicz and P. M. Marcus (1992). "Screening sigmoidoscopy and colorectal cancer mortality." Journal of the National Cancer Institute **84**(20): 1572-1575.

Nian, J., X. Sun, S. Ming, C. Yan, Y. Ma, Y. Feng, L. Yang, M. Yu, G. Zhang and X. Wang (2017). "Diagnostic Accuracy of Methylated SEPT9 for Blood-based Colorectal Cancer Detection: A Systematic Review and Meta-Analysis." Clin Transl Gastroenterol **8**(1): e216.

Niderman-Meyer, O., T. Zeidman, E. Shimoni and Y. Kashi (2010). "Mechanisms involved in governing adherence of Vibrio cholerae to granular starch." Appl Environ Microbiol **76**(4): 1034-1043.

Normand, S., A. Delanoye-Crespin, A. Bressenot, L. Huot, T. Grandjean, L. Peyrin-Biroulet, Y. Lemoine, D. Hot and M. Chamaillard (2011). "Nod-like receptor pyrin domain-containing protein 6 (NLRP6) controls epithelial self-renewal and colorectal carcinogenesis upon injury." Proceedings of the National Academy of Sciences **108**(23): 9601-9606.

Nugent, J. L., A. N. McCoy, C. J. Addamo, W. Jia, R. S. Sandler and T. O. Keku (2014). "Altered tissue metabolites correlate with microbial dysbiosis in colorectal adenomas." Journal of proteome research **13**(4): 1921-1929.

O'Hara, A. M. and F. Shanahan (2006). "The gut flora as a forgotten organ." EMBO reports **7**(7): 688-693.

Ohigashi, S., K. Sudo, D. Kobayashi, O. Takahashi, T. Takahashi, T. Asahara, K. Nomoto and H. Onodera (2013). "Changes of the intestinal microbiota, short chain fatty acids, and fecal pH in patients with colorectal cancer." Digestive diseases and sciences **58**(6): 1717-1726.

Ohigashi, S., K. Sudo, D. Kobayashi, T. Takahashi, K. Nomoto and H. Onodera (2013). "Significant changes in the intestinal environment after surgery in patients with colorectal cancer." Journal of Gastrointestinal Surgery **17**(9): 1657-1664.

Ohkusa, T., N. Sato, T. Ogihara, K. Morita, M. Ogawa and I. Okayasu (2002). "Fusobacterium varium localized in the colonic mucosa of patients with ulcerative colitis stimulates species-specific antibody." Journal of gastroenterology and hepatology **17**(8): 849-853.

Olsen, G. J., R. Overbeek, N. Larsen, T. L. Marsh, M. J. McCaughey, M. A. Maciukenas, W.-M. Kuan, T. J. Macke, Y. Xing and C. R. Woese (1992). "The ribosomal database project." Nucleic Acids Research **20**(suppl): 2199-2200.

Oswald, E. S., L. M. Brown, J. C. Bulinski and C. T. Hung (2011). "Label-free protein profiling of adipose-derived human stem cells under hyperosmotic treatment." Journal of proteome research **10**(7): 3050-3059.

Oulas, A., C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis and I. Iliopoulos (2015). "Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies." Bioinformatics and biology insights **9**: 75.

Park, D. I., S. Ryu, Y. H. Kim, S. H. Lee, C. K. Lee, C. S. Eun and D. S. Han (2010). "Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at average risk undergoing colorectal cancer screening." Am J Gastroenterol **105**(9): 2017-2025.

Parra-Blanco, A., A. Z. Gimeno-Garcia, E. Quintero, D. Nicolas, S. G. Moreno, A. Jimenez, M. Hernandez-Guerra, M. Carrillo-Palau, Y. Eishi and J. Lopez-Bastida (2010). "Diagnostic accuracy of immunochemical versus guaiac faecal occult blood tests for colorectal cancer screening." J Gastroenterol **45**(7): 703-712.

Peters, B. A., C. Dominianni, J. A. Shapiro, T. R. Church, J. Wu, G. Miller, E. Yuen, H. Freiman, I. Lustbader and J. Salik (2016). "The gut microbiota in conventional and serrated precursors of colorectal cancer." Microbiome **4**(1): 69.

Preheim, S. P., A. R. Perrotta, J. Friedman, C. Smilie, I. Brito, M. B. Smith and E. Alm (2013). Computational Methods for High-Throughput Comparative Analyses of Natural

Microbial Communities. Methods in Enzymology. J. N. Abelson and M. I. Simon. San Diego, CA, USA, Elsevier**:** 354-368.

Preheim, S. P., A. R. Perrotta, A. M. Martin-Platero, A. Gupta and E. J. Alm (2013). "Distribution-based clustering: using ecology to refine the operational taxonomic unit." Applied and environmental microbiology **79**(21): 6593-6603.

Price, M. N., P. S. Dehal and A. P. Arkin (2009). "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix." Molecular biology and evolution **26**(7): 1641-1650.

Price, M. N., P. S. Dehal and A. P. Arkin (2010). "FastTree 2–approximately maximum-likelihood trees for large alignments." PloS one **5**(3): e9490.

Proctor, L. M. (2011). "The human microbiome project in 2011 and beyond." Cell host & microbe **10**(4): 287-291.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies and F. O. Glöckner (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic acids research **35**(21): 7188-7196.

Pryde, S. E., S. H. Duncan, G. L. Hold, C. S. Stewart and H. J. Flint (2002). "The microbiology of butyrate formation in the human colon." FEMS microbiology letters **217**(2): 133-139.

Qian, B. Z. and J. W. Pollard (2010). "Macrophage diversity enhances tumor progression and metastasis." Cell **141**(1): 39-51.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H.

Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, H. I. T. C. Meta, P. Bork, S. D. Ehrlich and J. Wang (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan and D. Shen (2012). "A metagenome-wide association study of gut microbiota in type 2 diabetes." Nature **490**(7418): 55-60.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies and F. O. Glockner (2013). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." Nucleic Acids Res **41**(Database issue): D590-596.

Quince, C., A. Lanzen, R. J. Davenport and P. J. Turnbaugh (2011). "Removing noise from pyrosequenced amplicons." BMC bioinformatics **12**(1): 38.

Rajilic-Stojanovic, M. and W. M. de Vos (2014). "The first 1000 cultured species of the human gastrointestinal microbiota." FEMS Microbiol Rev **38**(5): 996-1047.

Ramakrishna, B. S., S. Venkataraman, P. Srinivasan, P. Dash, G. P. Young and H. J. Binder (2000). "Amylase-resistant starch plus oral rehydration solution for cholera." N Engl J Med **342**(5): 308-313.

Reeder, J. and R. Knight (2010). "Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions." Nature methods **7**(9): 668-669.

Reid, G. (2004). "When microbe meets human." Clin Infect Dis **39**(6): 827-830.

Renehan, A. G., M. Tyson, M. Egger, R. F. Heller and M. Zwahlen (2008). "Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies." The Lancet **371**(9612): 569-578.

Rideout, J. R., Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez and A. Robbins-Pianka (2014). "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences." PeerJ **2**: e545.

Rizzo, J. M. and M. J. Buck (2012). "Key principles and clinical applications of "next-generation" DNA sequencing." Cancer Prevention Research **5**(7): 887-900.

Robertson, M. D., A. S. Bickerton, A. L. Dennis, H. Vidal and K. N. Frayn (2005). "Insulin-sensitizing effects of dietary resistant starch and effects on skeletal muscle and adipose tissue metabolism." Am J Clin Nutr **82**(3): 559-567.

Rockey, D. C. (1999). "Occult gastrointestinal bleeding." New England Journal of Medicine **341**(1): 38-46.

Rodriguez, J. M., K. Murphy, C. Stanton, R. P. Ross, O. I. Kober, N. Juge, E. Avershina, K. Rudi, A. Narbad, M. C. Jenmalm, J. R. Marchesi and M. C. Collado (2015). "The composition of the gut microbiota throughout life, with an emphasis on early life." Microb Ecol Health Dis **26**: 26050.

Ronaghi, M. (2001). "Pyrosequencing sheds light on DNA sequencing." Genome research **11**(1): 3-11.

Rosen, G. L., E. R. Reichenberger and A. M. Rosenfeld (2011). "NBC: the Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads." Bioinformatics **27**(1): 127-129.

Rubinstein, M. R., X. Wang, W. Liu, Y. Hao, G. Cai and Y. W. Han (2013). "Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin." Cell host & microbe **14**(2): 195-206.

Russell, S. L., M. J. Gold, M. Hartmann, B. P. Willing, L. Thorson, M. Wlodarska, N. Gill, M. R. Blanchet, W. W. Mohn and K. M. McNagny (2012). "Early life antibiotic-driven

changes in microbiota enhance susceptibility to allergic asthma." EMBO reports **13**(5): 440-447.

Salminen, S., C. Bouley, M. C. Boutron-Ruault, J. H. Cummings, A. Franck, G. R. Gibson, E. Isolauri, M. C. Moreau, M. Roberfroid and I. Rowland (1998). "Functional food science and gastrointestinal physiology and function." Br J Nutr **80 Suppl 1**: S147-171.

Samad, A. K. A., R. S. Taylor, T. Marshall and M. A. S. Chapman (2005). "A meta-analysis of the association of physical activity with reduced risk of colorectal cancer." Colorectal Disease **7**(3): 204-213.

Sanapareddy, N., R. M. Legge, B. Jovov, A. McCoy, L. Burcal, F. Araujo-Perez, T. A. Randall, J. Galanko, A. Benson and R. S. Sandler (2012). "Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans." The ISME journal **6**(10): 1858-1868.

Savage, D. C. (1977). "Microbial ecology of the gastrointestinal tract." Annu Rev Microbiol **31**: 107-133.

Sayad, S. (2011). Real Time Data Mining, Self-Help Publishers.

Sboner, A., X. J. Mu, D. Greenbaum, R. K. Auerbach and M. B. Gerstein (2011). "The real cost of sequencing: higher than you think!" Genome biology **12**(8): 125.

Schloss, P. D. and J. Handelsman (2005). "Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness." Applied and environmental microbiology **71**(3): 1501-1506.

Schloss, P. D. and S. L. Westcott (2011). "Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis." Applied and environmental microbiology **77**(10): 3219-3226.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks and C. J. Robinson (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." Applied and environmental microbiology **75**(23): 7537-7541.

Schubert, A. M., H. Sinani and P. D. Schloss (2015). "Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against Clostridium difficile." MBio **6**(4): e00974-00915.

Schwabe, R. F. and C. Jobin (2013). "The microbiome and cancer." Nature Reviews Cancer **13**(11): 800-812.

Sears, C. L. (2009). "Enterotoxigenic Bacteroides fragilis: a rogue among symbiotes." Clin Microbiol Rev **22**(2): 349-369, Table of Contents.

Sears, C. L. and W. S. Garrett (2014). "Microbes, microbiota, and colon cancer." Cell host & microbe **15**(3): 317-328.

Segata, N., J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett and C. Huttenhower (2011). "Metagenomic biomarker discovery and explanation." Genome Biol **12**(6): R60.

Selby, J. V., G. D. Friedman, C. P. Quesenberry Jr and N. S. Weiss (1992). "A case–control study of screening sigmoidoscopy and mortality from colorectal cancer." New England Journal of Medicine **326**(10): 653-657.

Shaukat, A., A. Dostal, J. Menk and T. R. Church (2017). "BMI Is a Risk Factor for Colorectal Cancer Mortality." Dig Dis Sci.

Shen, X. J., J. F. Rawls, T. A. Randall, L. Burcall, C. Mpande, N. Jenkins, B. Jovov, Z. Abdo, R. S. Sandler and T. O. Keku (2010). "Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas." Gut microbes **1**(3): 138-147.

Shiryaev, S. A., A. G. Remacle, A. V. Chernov, V. S. Golubkov, K. Motamedchaboki, N. Muranaka, C. M. Dambacher, P. Capek, M. Kukreja, I. A. Kozlov, M. Perucho, P. Cieplak and A. Y. Strongin (2013). "Substrate cleavage profiling suggests a distinct function of Bacteroides fragilis metalloproteinases (fragilysin and metalloproteinase II) at the microbiome-inflammation-cancer interface." J Biol Chem **288**(48): 34956-34967.

Siegel, R. L., K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi and A. Jemal (2017). "Colorectal cancer statistics, 2017." CA Cancer J Clin **67**(3): 177-193.

Siegel, R. L., K. D. Miller and A. Jemal (2015). "Cancer statistics, 2015." CA: a cancer journal for clinicians **65**(1): 5-29.

Simon, R., M. D. Radmacher, K. Dobbin and L. M. McShane (2003). "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." Journal of the National Cancer Institute **95**(1): 14-18.

Smith, M. I., T. Yatsunenko, M. J. Manary, I. Trehan, R. Mkakosya, J. Cheng, A. L. Kau, S. S. Rich, P. Concannon and J. C. Mychaleckyj (2013). "Gut microbiomes of Malawian twin pairs discordant for kwashiorkor." Science **339**(6119): 548-554.

Sohn, D. K., S. Y. Jeong, H. S. Choi, S. B. Lim, J. M. Huh, D. H. Kim, D. Y. Kim, Y. H. Kim, H. J. Chang, K. H. Jung, J. B. Ahn, H. K. Kim and J. G. Park (2005). "Single immunochemical fecal occult blood test for detection of colorectal neoplasia." Cancer Res Treat **37**(1): 20-23.

Sokol, S. Y. (1999). "Wnt signaling and dorso-ventral axis specification in vertebrates." Curr Opin Genet Dev **9**(4): 405-410.

Sonnenburg, J. L., L. T. Angenent and J. I. Gordon (2004). "Getting a grip on things: how do communities of bacterial symbionts become established in our intestine?" Nat Immunol **5**(6): 569-573.

Spencer, M. D., T. J. Hamp, R. W. Reid, L. M. Fischer, S. H. Zeisel and A. A. Fodor (2011). "Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency." Gastroenterology **140**(3): 976-986.

Stark, M., S. A. Berger, A. Stamatakis and C. von Mering (2010). "MLTreeMap-accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies." BMC genomics **11**(1): 461.

Stewart, S. L., J. M. Wike, I. Kato, D. R. Lewis and F. Michaud (2006). "A population-based study of colorectal cancer histology in the United States, 1998–2001." Cancer **107**(S5): 1128-1141.

Strauss, J., G. G. Kaplan, P. L. Beck, K. Rioux, R. Panaccione, R. DeVinney, T. Lynch and E. Allen-Vercoe (2011). "Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host." Inflammatory bowel diseases **17**(9): 1971-1978.

Strum, W. B. (2016). "Colorectal adenomas." New England Journal of Medicine **374**(11): 1065-1075.

Sun, Y., Y. Cai, S. M. Huse, R. Knight, W. G. Farmerie, X. Wang and V. Mai (2012). "A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis." Brief Bioinform **13**(1): 107-121.

Sun, Y., Y. Cai, L. Liu, F. Yu, M. L. Farrell, W. McKendree and W. Farmerie (2009). "ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences." Nucleic acids research **37**(10): e76.

Swidsinski, A., Y. Dörffel, V. Loening-Baucke, F. Theissig, J. C. Rückert, M. Ismail, W. A. Rau, D. Gaschler, M. Weizenegger and S. Kühn (2009). "Acute appendicitis is characterized by local invasion with Fusobacterium nucleatum/necrophorum." Gut: gut. 2009.191320.

Tamanai-Shacoori, Z., I. Smida, L. Bousarghin, O. Loreal, V. Meuric, S. B. Fong, M. Bonnaure-Mallet and A. Jolivet-Gougeon (2017). "Roseburia spp.: a marker of health?" Future Microbiol **12**: 157-170.

Tan, A. C., D. Q. Naiman, L. Xu, R. L. Winslow and D. Geman (2005). "Simple decision rules for classifying human cancers from gene expression profiles." Bioinformatics **21**(20): 3896-3904.

Tan, P.-N., M. Steinbach and V. Kumar (2006). Introduction to data mining. Library of Congress.

Taylor, D. P., R. W. Burt, M. S. Williams, P. J. Haug and L. A. Cannon–Albright (2010). "Population-based family history–specific risks for colorectal cancer: a constellation approach." Gastroenterology **138**(3): 877-885.

Teeling, H., J. Waldmann, T. Lombardot, M. Bauer and F. O. Glöckner (2004). "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences." BMC bioinformatics **5**(1): 163.

Thomas, A. M., E. C. Jesus, A. Lopes, S. Aguiar, Jr., M. D. Begnami, R. M. Rocha, P. A. Carpinetti, A. A. Camargo, C. Hoffmann, H. C. Freitas, I. T. Silva, D. N. Nunes, J. C. Setubal and E. Dias-Neto (2016). "Tissue-Associated Bacterial Alterations in Rectal Carcinoma Patients Revealed by 16S rRNA Community Profiling." Front Cell Infect Microbiol **6**: 179.

Tinmouth, J., I. Lansdorp-Vogelaar and J. E. Allison (2015). "Faecal immunochemical tests versus guaiac faecal occult blood tests: what clinicians and colorectal cancer screening programme organisers need to know." Gut **64**(8): 1327-1337.

Tosolini, M., A. Kirilovsky, B. Mlecnik, T. Fredriksen, S. Mauger, G. Bindea, A. Berger, P. Bruneval, W. H. Fridman, F. Pages and J. Galon (2011). "Clinical impact of different classes of infiltrating T cytotoxic and helper cells (Th1, th2, treg, th17) in patients with colorectal cancer." Cancer Res **71**(4): 1263-1271.

Toth, K., R. Wasserkort, F. Sipos, A. Kalmar, B. Wichmann, K. Leiszter, G. Valcz, M. Juhasz, P. Miheller, A. V. Patai, Z. Tulassay and B. Molnar (2014). "Detection of methylated septin 9 in tissue and plasma of colorectal patients with neoplasia and the relationship to the amount of circulating cell-free DNA." PLoS One **9**(12): e115415.

Tringe, S. G. and P. Hugenholtz (2008). "A renaissance for the pioneering 16S rRNA gene." Current opinion in microbiology **11**(5): 442-446.

Tringe, S. G., C. Von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur and J. C. Detter (2005). "Comparative metagenomics of microbial communities." Science **308**(5721): 554-557.

Turnbaugh, P. J., F. Bäckhed, L. Fulton and J. I. Gordon (2008). "Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome." Cell host & microbe **3**(4): 213-223.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe and J. P. Affourtit (2009). "A core gut microbiome in obese and lean twins." nature **457**(7228): 480-484.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis and J. I. Gordon (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**(7122): 1027-1031.

Vaahtovuo, J., E. Munukka, M. KorkeamÄKi, R. Luukkainen and P. Toivanen (2008). "Fecal microbiota in early rheumatoid arthritis." The Journal of rheumatology **35**(8): 1500-1505.

Van de Peer, Y., S. Chapelle and R. De Wachter (1996). "A quantitative map of nucleotide substitution rates in bacterial rRNA." Nucleic acids research **24**(17): 3381-3391.

Van den Abbeele, P., C. Belzer, M. Goossens, M. Kleerebezem, W. M. De Vos, O. Thas, R. De Weirdt, F. M. Kerckhof and T. Van de Wiele (2013). "Butyrate-producing

Clostridium cluster XIVa species specifically colonize mucins in an in vitro gut model." ISME J **7**(5): 949-961.

Vital, M., A. C. Howe and J. M. Tiedje (2014). "Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data." MBio **5**(2): e00889.

Vitetta, L., D. Briskey, H. Alford, S. Hall and S. Coulson (2014). "Probiotics, prebiotics and the gastrointestinal tract in health and disease." Inflammopharmacology **22**(3): 135-154.

Vogelstein, B. and K. W. Kinzler (1993). "The multistep nature of cancer." Trends in genetics **9**(4): 138-141.

Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz and K. W. Kinzler (2013). "Cancer genome landscapes." science **339**(6127): 1546-1558.

Walker, A. W., J. Ince, S. H. Duncan, L. M. Webster, G. Holtrop, X. Ze, D. Brown, M. D. Stares, P. Scott, A. Bergerat, P. Louis, F. McIntosh, A. M. Johnstone, G. E. Lobley, J. Parkhill and H. J. Flint (2011). "Dominant and diet-responsive groups of bacteria within the human colonic microbiota." ISME J **5**(2): 220-230.

Walters, W. A., Z. Xu and R. Knight (2014). "Meta-analyses of human gut microbes associated with obesity and IBD." FEBS Lett **588**(22): 4223-4233.

Wang, L., C. T. Christophersen, M. J. Sorich, J. P. Gerber, M. T. Angley and M. A. Conlon (2013). "Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder." Mol Autism **4**(1): 42.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and environmental microbiology **73**(16): 5261-5267.

Wang, X., Y. Yang and M. M. Huycke (2017). "Microbiome-driven carcinogenesis in colorectal cancer: Models and mechanisms." Free Radical Biology and Medicine **105**: 3-15.

Wei, H., L. Dong, T. Wang, M. Zhang, W. Hua, C. Zhang, X. Pang, M. Chen, M. Su and Y. Qiu (2010). "Structural shifts of gut microbiota as surrogate endpoints for monitoring host health changes induced by carcinogen exposure." FEMS microbiology ecology **73**(3): 577-586.

Weir, T. L., D. K. Manter, A. M. Sheflin, B. A. Barnett, A. L. Heuberger and E. P. Ryan (2013). "Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults." PLoS One **8**(8): e70803.

Weirather, J. L., M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X. J. Wang, D. Buck and K. F. Au (2017). "Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis." F1000Res **6**: 100.

Wexler, H. (2005). Genus VIII. Sutterella, p 682–683. Brenner DJ, Krieg NR, Staley JT (ed), Bergey's manual of systematic bacteriology, vol 2, part C. The Proteobacteria, Springer-Verlag, New York, NY.

Wexler, H. M. (2007). "Bacteroides: the good, the bad, and the nitty-gritty." Clin Microbiol Rev **20**(4): 593-621.

White, J. R., N. Nagarajan and M. Pop (2009). "Statistical methods for detecting differentially abundant features in clinical metagenomic samples." PLoS computational biology **5**(4): e1000352.

Win, A. K., R. J. MacInnis, J. L. Hopper and M. A. Jenkins (2012). "Risk prediction models for colorectal cancer: a review." Cancer Epidemiology Biomarkers & Prevention **21**(3): 398-410.

Wisittipanit, N., H. Rangwala, M. Sikaroodi, A. Keshavarzian, E. A. Mutlu and P. Gillevet (2015). "Classification methods for the analysis of LH–PCR data associated with inflammatory bowel disease patients." International journal of bioinformatics research and applications **11**(2): 111-129.

Witten, I. H. and E. Frank (2005). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

Woese, C. R. (1987). "Bacterial evolution." Microbiological reviews **51**(2): 221.

Wong, S. H., L. Zhao, X. Zhang, G. Nakatsu, J. Han, W. Xu, X. Xiao, T. N. Kwong, H. Tsoi, W. K. Wu, Z. Benhua, F. K. Chan, J. J. Sung, H. Wei and J. Yu (2017). "Gavage of Fecal Samples From Patients with Colorectal Cancer Promotes Intestinal Carcinogenesis in Germ-free and Conventional Mice." Gastroenterology.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu and B. J. Tindall (2009). "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea." Nature **462**(7276): 1056-1060.

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters and R. Knight (2011). "Linking long-term dietary patterns with gut microbial enterotypes." Science **334**(6052): 105-108.

Wu, N., X. Yang, R. Zhang, J. Li, X. Xiao, Y. Hu, Y. Chen, F. Yang, N. Lu and Z. Wang (2013). "Dysbiosis signature of fecal microbiota in colorectal cancer patients." Microbial ecology **66**(2): 462-470.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu and S. Y. Philip (2008). "Top 10 algorithms in data mining." Knowledge and Information Systems **14**(1): 1-37.

Xu, K. and B. Jiang (2017). "Analysis of Mucosa-Associated Microbiota in Colorectal Cancer." Med Sci Monit **23**: 4422-4430.

Yamaoka, Y., Y. Suehiro, S. Hashimoto, T. Hoshida, M. Fujimoto, M. Watanabe, D. Imanaga, K. Sakai, T. Matsumoto and M. Nishioka (2017). "Fusobacterium nucleatum as a prognostic marker of colorectal cancer in a Japanese population." Journal of Gastroenterology: 1-8.

Yang, X. O., S. H. Chang, H. Park, R. Nurieva, B. Shah, L. Acero, Y.-H. Wang, K. S. Schluns, R. R. Broaddus and Z. Zhu (2008). "Regulation of inflammatory responses by IL-17F." Journal of Experimental Medicine **205**(5): 1063-1075.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano and A. P. Anokhin (2012). "Human gut microbiome viewed across age and geography." Nature **486**(7402): 222-227.

Yoon, H., N. Kim, J. H. Park, Y. S. Kim, J. Lee, H. W. Kim, Y. J. Choi, C. M. Shin, Y. S. Park, D. H. Lee and H. C. Jung (2017). "Comparisons of Gut Microbiota Among Healthy Control, Patients With Conventional Adenoma, Sessile Serrated Adenoma, and Colorectal Cancer." J Cancer Prev **22**(2): 108-114.

Young, G. P., Y. Hu, R. K. Le Leu and L. Nyskohus (2005). "Dietary fibre and colorectal cancer: a model for environment--gene interactions." Mol Nutr Food Res **49**(6): 571-584.

Zackular, J. P., N. T. Baxter, G. Y. Chen and P. D. Schloss (2015). "Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis." mSphere **1**(1).

Zackular, J. P., N. T. Baxter, K. D. Iverson, W. D. Sadler, J. F. Petrosino, G. Y. Chen and P. D. Schloss (2013). "The gut microbiome modulates colon tumorigenesis." MBio **4**(6): e00692-00613.

Zackular, J. P., M. A. M. Rogers, M. T. Ruffin and P. D. Schloss (2014). "The human gut microbiome as a screening tool for colorectal cancer." Cancer Prevention Research **7**(11): 1112-1121.

Zauber, A. G., S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, W. Shi, J. H. Bond, M. Schapiro and J. F. Panish (2012). "Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths." New England Journal of Medicine **366**(8): 687-696.

Ze, X., S. H. Duncan, P. Louis and H. J. Flint (2012). "Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon." The ISME journal **6**(8): 1535-1543.

Zhang, C. and M. Zhong (2015). "Consumption of beer and colorectal cancer incidence: a meta-analysis of observational studies." Cancer Causes & Control **26**(4): 549-560.

Zheng, H. and H. Wu (2010). "Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis." Journal of bioinformatics and computational biology **8**(06): 995-1011.

Zhu, L., S. S. Baker, C. Gill, W. Liu, R. Alkhouri, R. D. Baker and S. R. Gill (2013). "Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH." Hepatology **57**(2): 601-609.

BIOGRAPHY

Ezzat Dadkhah graduated from Farzanegan High School, Mashhad, Iran, in 1997. She received her Bachelor of Science in Microbiology from Alzahra University, Tehran, Iran in 2003 and her Master of Science in Cell & Molecular Biology from Islamic Azad University in 2009, Tehran, Iran. She was employed as a research associate in Medical Genetics Division, Mashhad University of Medical Sciences, Mashhad, Iran for eight years. She also worked as teaching assistant in Biology Department, George Mason University, Fairfax, VA for four years.