Programa Interunidades de Pós-Graduação em Bioinformática
Universidade de São Paulo

Microbial community profiling of human gastrointestinal cancers

Andrew Maltez Thomas

PhD thesis

Supervisor: João Carlos Setubal
Co-supervisor: Emmanuel Dias-Neto

São Paulo
2018

FICHA CATALOGRÁFICA

Thomas, Andrew Maltez
T454 Microbial community profiling of human cancers / Andrew Maltez Thomas João Carlos Setubal, [orient.] e Emmanuel Dias-Neto, [coorient]. São Paulo : 2018.
91 p.

Tese (Doutorado) - Universidade de São Paulo
Orientador: Prof. Setubal, João Carlos
Coorientador: Prof. Dias-Neto, Emmanuel

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Bioinformática. 2. Neoplasias 3. Ecologia de Comunidades. 4. Genética Microbiana. I. Setubal, João Carlos, orientador. II. Dias-Neto, Emmanuel, coorient. III. Universidade de São Paulo. IV. Título.

CDD – 572.8

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela bibliotecária Maria Lucia Ribeiro CRB-8/ 2766

Acknowledgments

***To my parents, Patricia and Stephen**, for all of their support and understanding during this period. Without you this wouldn't be possible.*

***To Daiana Marin**, for the unconditional love, support, understanding and encouragement.*

***To my extended family (grandparents, uncles, aunts and cousins)**, for all the joy and positive energy you bring.*

***To my supervisor, Prof. Dr. João Carlos Setubal**, for the opportunity, guidance and liberty during this work.*

***To my co-supervisor, Prof. Dr. Emmanuel Dias-Neto**, for all the support, guidance and liberty during this work.*

***To Prof. Dr. Nicola Segata**, for the opportunity, support and guidance during the scientific internship.*

***To Dr. Diana Noronha Nunes**, for all her help and advice.*

***To Dr. Eliane de Camargo Jesus**, for providing a great partnership and aiding with sample collection.*

***To the members of the Bioinformatics Laboratory**, for our scientific exchanges and laid back moments.*

***To the members of the Medical Genomics Laboratory**, Thais, Gabriela Albuquerque, Marianna, Marina, Emilio, Maria, Gabriela Branco, Luiza, Helano, Jordana and Isabella for all their help.*

***To the members of the Computational Metagenomics Laboratory, especially Paolo Manghi**, for your friendship and partnership.*

***To Lais Senda and Dr. Adriane Pelosof**, for their help with sample collection.*

***To Drs Ademar Lopes and Samuel Aguiar Junior**, for their help.*

***To Dr. Rafael Malagoli Rocha**, for his help with the immunohistochemistry analysis.*

***To Drs Anamaria Aranha Camargo and Paola Avelar Carpinetti**, for their help with ddPCR.*

***To Prof. Dr. Levi Waldron**, for his help with the statistical analysis.*

***To Drs Alessio Naccarati, Barbara Pardini, Francesca Cordero, Sonia Tarallo, Antonio Francavilla, Gaetano Gallo, Mario Trompetto and Giulio Ferrero,** for their help with sample collection and analysis.*

To Drs Maria Rescigno, Chiara Pozzi, Sara Gandini and Davide Serrano, *for their help with sample collection.*

To Israel Tojal and Rodrigo Borges*, for their help with the bioinformatic analysis.*

To the members of the examining board*.*

To all patients, *who provided samples for this thesis.*

To CAPES*, for their financial support.*

To FAPESP, *for the doctoral scholarship (15/01507-7), the BEPE scholarship (16/23527-2) and funding for the experimental work (14/26897-0).*

To PRONON*, for the financial support (SIPAR 2500.055-167/2015-23).*

To AC Camargo Cancer Center**,** for providing access to clinical data, the infrastructure for sample collection, storage and experimental work.

To the Chemistry Institute - University of São Paulo**,** for providing the infrastructure for data storage and analysis.

Abstract

The human microbiome - defined as the microbial communities that live in and on our bodies - is emerging as a key factor in human diseases. The expanding research field that investigates the role of the microbiome on human cancer development, termed oncobiome, has led to important discoveries such as the role of *Fusobacterium nucleatum* in colorectal cancer carcinogenesis and tumor progression. Motivated by these discoveries, this thesis studied the oncobiome from different perspectives, investigating whether alterations to microbial profiles were associated with disease status or an adverse response to treatment. We used both biopsy tissue samples and 16S rRNA amplicon sequencing (N = 36), as well as privately and publicly available fecal whole metagenomes (N = 764) to investigate microbiome-colorectal cancer (CRC) associations. We observed significant increases in species richness in CRC, regardless of sample type or methodology, which was partially due to expansions of species typically from the oral cavity, as well as an overabundance of specific taxa such as *Bacteroides fragilis*, *Fusobacterium*, *Desulfovibrio* and *Bilophila* in CRC. Functional potential analysis of CRC metagenomes revealed that the choline trimethylamine-lyase (*cutC*) gene was over-abundant in CRC, with the strength of association dependent on four identified sequence variants, pointing at a novel potential mechanism of CRC carcinogenesis. Predictive microbiome signatures trained on the combination of multiple datasets showed very high and consistent performances on distinct cohorts (average AUC 0.83, minimum 0.81). To investigate the microbiome's role in response to treatment, we profiled microbial communities of gastric wash samples in gastric cancer patients (N = 36) before and after neoadjuvant chemotherapy through 16S rRNA amplicon sequencing. Gastric wash microbial communities presented remarkably high inter-individual variation, with significant decreases in richness and phylogenetic diversity after treatment and associations with pH, pathological response and sample collection. The most abundant genera found in patients before or after chemotherapy treatment included *Streptococcus*, *Prevotella*, *Rothia* and *Veillonella*. Despite limitations inherent to differing experimental choices, this thesis provides microbiome signatures that can be the basis for clinical prognostic tests and hypothesis-driven mechanistic studies, as well as supporting the role of the human oral microbiome in whole-body diseases.

Keywords: Oncobiome; metagenomics; 16s rRNA; colorectal cancer; gastric cancer; microbiome

Resumo

O microbioma humano - definido como as comunidades microbianas que vivem sobre e dentro do corpo humano - está se tornando um fator cada vez mais importante em doenças humanas. O campo de estudo que investiga o papel do microbioma no desenvolvimento do câncer humano, denominado *oncobioma*, está crescendo e já levou a importantes descobertas como o papel da espécie *Fusobacterium nucleatum* na carcinogênese e progressão tumoral de tumores colorretais. Motivado por estas descobertas, esta tese de doutorado analisou o oncobioma por diferentes perspectivas, investigando se alterações nos perfis microbianos estavam associados à presença da doença ou a uma resposta adversa ao tratamento. Usamos tanto amostras de tecidos de biópsias e o sequenciamento do gene 16S rRNA (N = 36), quanto metagenomas fecais públicos e privados (N = 764), para investigar associações entre o microbioma e o câncer colorretal (CCR). Observamos um aumento significativo da riqueza microbiana no CCR, independentemente do tipo da amostra ou metodologia, que era em parte, devido ao aumento de espécies tipicamente presentes na cavidade oral. Observamos também um aumento da abundância de táxons específicos no CCR, que incluíam *Bacteroides fragilis*, *Fusobacterium*, *Desulfovibrio* e *Bilophila*. Analisando o potencial funcional dos metagenomas, encontramos um aumento significativo da enzima liase colina trimetilamina (*cutC*) no CCR, cuja associação era dependente de 4 variantes de sequência, demonstrando ser um possível novo mecanismo de carcinogênese no CCR. Assinaturas preditivas do microbioma treinadas na combinação dos estudos demonstraram ser altamente preditivas e consistentes nos diferentes estudos (média de AUC 0.83, mínimo de 0.81). Para investigar o possível papel do microbioma na resposta ao tratamento, analisamos os perfis microbianos do suco gástrico de pacientes com câncer gástrico (N = 36) antes e depois do tratamento quimioterápico neoadjuvante. As comunidades microbianas apresentaram uma variabilidade inter-individual notavelmente grande, com diminuições significativas na riqueza e diversidade filogenética pós tratamento, além de estarem associadas principalmente ao pH, mas também à resposta patológica e ao tempo da coleta. Os gêneros mais abundantes encontrados nos pacientes antes ou depois da quimioterapia incluíam *Streptococcus*, *Prevotella*, *Rothia* e *Veillonella*. Apesar das limitações inerentes às escolhas experimentais, esta tese proporciona assinaturas do microbioma que podem servir de base para testes clínicos prognósticos e estudos mecanísticos, além de dar mais suporte ao papel do microbioma oral em doenças humanas.

Palavras-chave: oncobioma; metagenômica; 16s rRNA; câncer colorretal; câncer gástrica**;** microbioma

Table of contents

Abbreviations

16s rRNA	16s ribosomal ribonucleic acid
ASV	Amplicon sequence variant
ANOSIM	Analysis of similarities
AUC	Area under the receiver operator curve
BMI	Body mass index
CRC	Colorectal cancer
ddPCR	Digital droplet polymerase chain reaction
DNA	Deoxyribonucleic acid
EBV	Epstein-Barr virus
GC	Gastric cancer
HP	*Helicobacter pylori*
HTS	High throughput sequencing
IBD	Irritable bowel syndrome
PCR	Polymerase chain reaction
PPI	Proton pump inhibitor
NC	Non-cancer
nt	nucleotide
OTU	Operational taxonomic unit
OR	Odds ratio
qiime	Quantitative insights into microbial ecology
RC	Rectal-cancer
RDP	Ribosomal database project
TMA	Trimethylamine
SNV	Single nucleotide variant
WMS	Whole metagenome shotgun

Figures

Tables

Chapter 1. Introduction

1.1 The human gut microbiome

The human gut microbiome - defined as the collection of microbial genes, genomes and products that populate our intestinal tract - is the most diverse of the human body, with an estimated richness of 500-1000 bacterial species (Ramakrishna 2007) representing the bulk of the microbiota (10^{12} bacteria/gm feces) (Sommer and Bäckhed 2013) and about 10 million genes (Qin et al. 2010). The human gut microbiome maintains a symbiotic relationship with the gut mucosa and imparts substantial metabolic, immunological and gut protective functions in the healthy individual (Jandhyala et al. 2015). Strain population structures from the vast majority of species found in the human gut exhibit genomic variation, persisting within their hosts for years, indicating a stable co-existent ecosystem (Faith et al. 2013; Schloissnig et al. 2012). Strains are largely host-specific (Faith et al. 2013; Schloissnig et al. 2012), with similar strains being shared among closely related individuals. However, some strains of the same species may encode considerably different sets of genes and gene copy numbers (Greenblum et al. 2015), with up to 30% difference between their genomes (Zhang and Zhao 2016). Such intra-species variation endows each strain with potentially distinct functional capabilities, including virulence (Gill et al. 2005; Salama et al. 2000; Solheim et al. 2009), motility (Zunino et al. 1994), nutrient utilization (Siezen et al. 2010), and drug resistance (Gill et al. 2005), with trends identified in species profiles often being poorly translated to gene profiles and vice versa (Muegge et al. 2011; Turnbaugh et al. 2008).

The understanding of our microbiota, together with the determination of its composition when contrasting healthy vs. diseased states allows the identification of microorganism disturbances that are possibly related to disease development and offers a new approach for diagnosis as well as preventive and therapeutic interventions. Therefore, studies have investigated perturbations to the human gut microbiome related to disease, including inflammatory bowel disease (IBD), Crohn's disease, obesity, colorectal cancer, among others (Turnbaugh et al. 2008; Kostic et al. 2012; Gevers et al. 2014).

1.2 The oncobiome

In 2002, estimates indicated that infectious agents had a central role in 17.8% of all hepatic, gastric and cervical cancers (1.9 million cases) (Parkin 2006). Motivated by these estimates, the fact that several body sites (such as the oral cavity, gastrointestinal tract and skin) have continuous contact with the exterior world

and cancers genetics can only explain a small proportion of disease incidence, studies have focused on the interplay between the human microbiome and cancer development, known as the ‘oncobiome’ (Thomas and Jobin 2015). The oncogenic role of viruses, such as the Human Papillomavirus (HPV), has long been recognized (Moore and Chang 2010), however there is still little direct evidence that the symbiotic microbiota can modulate carcinogenesis in humans. The most recognized link between bacteria and cancer is the case of *Helicobacter pylori* (HP) and gastric adenocarcinomas (Marshall and Warren 1984; Peek and Crabtree 2006). This bacterium has been shown to secrete several virulence factors which include *CagA* (cytotoxin-associated gene A), *VacA* (vacuolating cytotoxin A), *urease*, and *NapA2* (neutrophil-activating protein A) that cause oxidative stress, chronic inflammation, and host DNA damage, leading to cancer (Hardbower et al. 2013; Koeppel et al. 2015; Wroblewski and Peek 2013). HP is the first bacterium to be designated a type I carcinogen by the World Health Organization.

Carcinogenesis is an inherently inflammatory process, with many proinflammatory and immunosuppressive pathways acting in conjunction with neoplastic processes. Microbial products, such as lipopolysaccharide (LPS), hydrogen sulfide and reactive oxygen species, can impact the innate and adaptive immune system and favor tumor growth in the colon (Grivennikov et al. 2012), or even provide mechanisms for tumor immune evasion (Gur et al. 2015). Bacteria-derived genotoxic substances, such as the toxin *fragilysin* produced by the Enterotoxigenic *Bacteroides fragilis* and the toxin *colibactin* produced by *Escherichia coli* of the B2 phylogenetic group, have the ability to cause DNA damage. So far, studies have identified roles of specific bacteria in carcinogenesis (Rubinstein et al. 2013; Kostic et al. 2012), modulation of the tumor microenvironment (Kostic et al. 2013), and interference with anti-cancer immune responses and immune-surveillance that facilitate chemotherapy activity (Zitvogel et al. 2013; Galluzzi et al. 2015; Vétizou et al. 2015) (**Figure 1**). The emerging concept that cancer needs to be studied considering the complex tumor microenvironment, which includes components such as tumor cells, immune infiltrates and the surrounding microenvironment and the microbiome, may aid in the development and improvement of cancer treatment, including immunotherapy (Pitt et al. 2016).

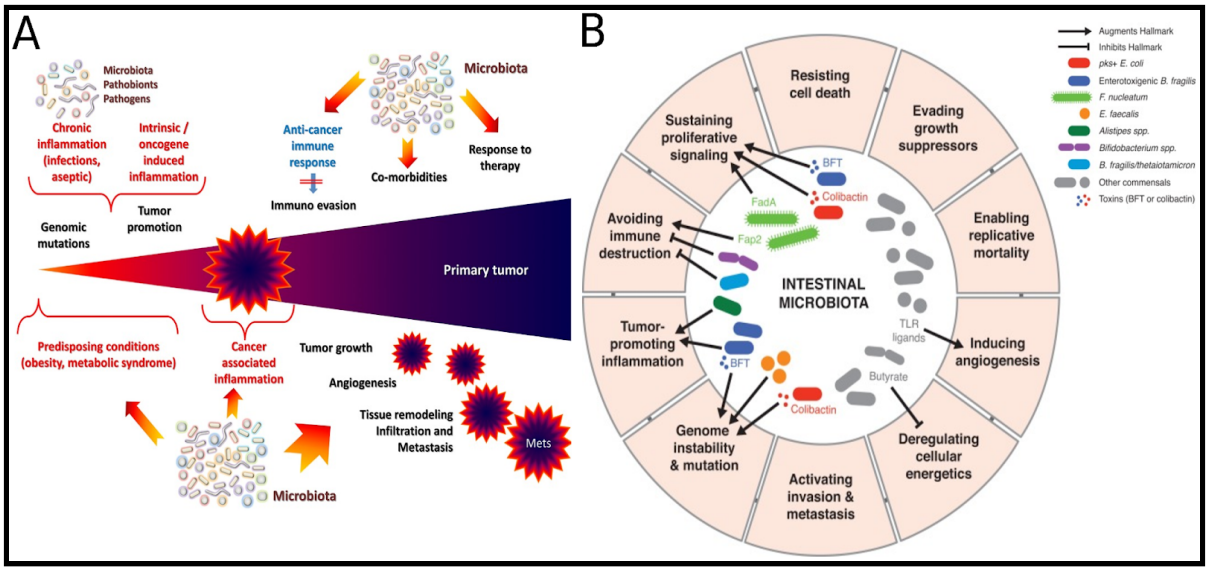


Figure 1. Role of the microbiota in cancer initiation, promotion, dissemination, and response to therapy. (a) The microbiota can directly and indirectly affect the development and maintenance of inflammation, which is one of the hallmarks of cancer. The microbiota can also affect the development of cancer-promoting conditions, such as obesity and metabolic syndrome, and modulate many of the inflammatory and immune mechanisms regulating cancer initiation and progression. **Source**: Dzutsev et al. 2015. **(b)** Recent advances in human microbiome research show mechanistic evidence of how the microbiome can modulate different hallmarks of cancer. **Source:** Fulbright et al. 2017.

1.3 Techniques for studying the microbiome

Culture-based studies have contributed significantly to our understanding of the microbial world, but over the years have reached technical limitations inherent to the methodology. Since about 1% of the microbial world is culturable with current techniques (Schloss and Handelsman 2005), this methodology has taken a back seat in recent years. A great leap forward in our understanding of the human microbiome came with a significant drop in the cost of DNA sequencing and a method termed “metagenomics”, whereby the whole genetic material of a microbial community is sequenced. Microbial community membership and composition is then determined through bioinformatic and ecological approaches (**Figure 2**).

There are two major strategies for analysing metagenomic DNA; via polymerase chain reaction (PCR) amplification of the prokaryotic 16S rRNA gene or via whole metagenome shotgun (WMS) sequencing. Each strategy has its pros and cons, depending on the aims of the study and tissue-type, which impacts the human/non-human nucleic acids ratio. 16S rRNA amplicon sequencing is quick, simple and with relatively inexpensive sample preparation and bioinformatic infrastructure and analysis. It is amenable to low-biomass and highly host-contaminated samples, providing a snapshot of community composition, membership and structure. However, it is subject to amplification biases (Bonnet et al. 2002), which can be magnified by the choice of primers and variable region (Walker et al. 2015), that in most cases leads to limited genus level resolution and

functional information. WMS sequencing on the other hand can directly infer the relative abundance of microbial functional genes, provide species and strain-level resolution (Scholz et al. 2016), capture phages, viruses, plasmids, microbial eukaryotes and does not suffer from PCR-related biases. It can also estimate *in situ* growth rates for target organisms with sequenced genomes (Korem et al. 2015), allows the recovery of genomes (metagenomic assembled genomes - MAGs) (Parks et al. 2017) and can be used to mine for novel gene families. However, this method is relatively expensive, laborious and with complex bioinformatic infrastructure and analysis. Samples with a high contamination of host-derived DNA and organelles may obscure microbial signatures and large sequencing depths are typically required. Other techniques to study the microbiome include metatranscriptomics, metametabolomics, metaproteomics and microarrays. However, some of these methods are still emerging and without comprehensive databases or analytical pipelines, have not yet been widely adopted.

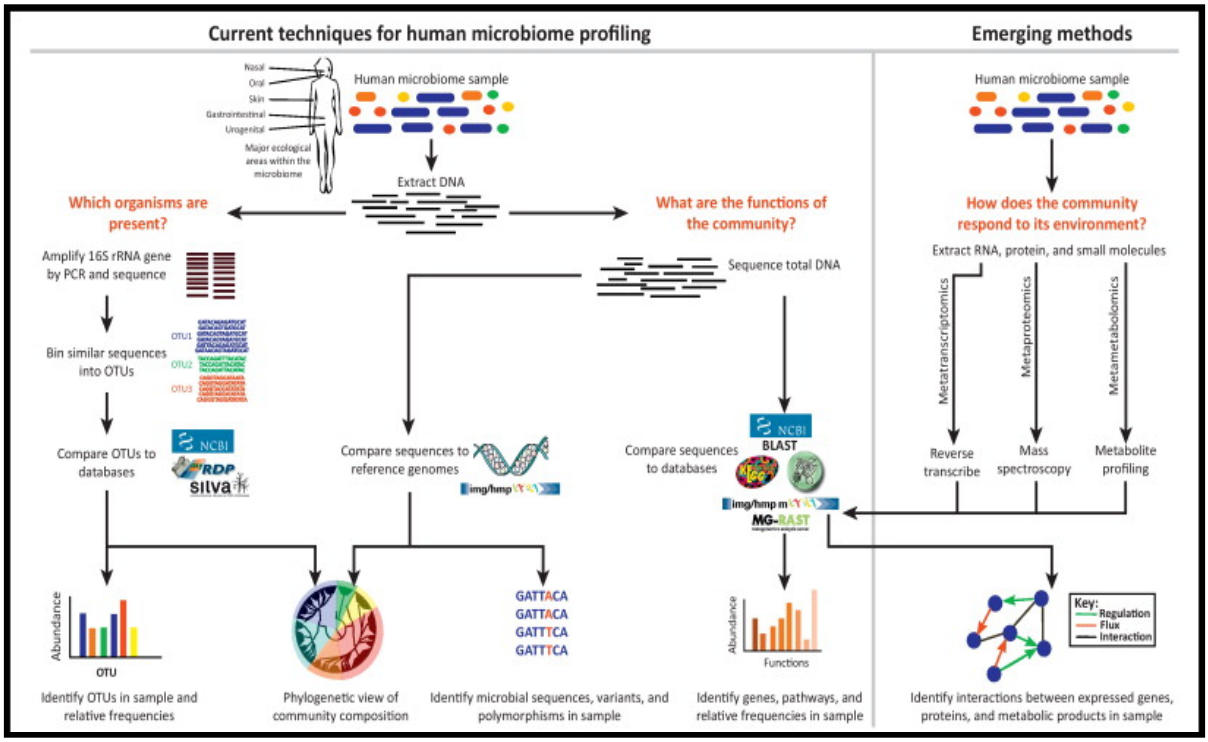


Figure 2. Current and emerging bioinformatic methods for studying the human microbiome. Community DNA is extracted from a sample containing several microbial species (colored differently). Total DNA is extracted and the composition of the community can be determined by amplifying and sequencing the 16S rRNA gene. Highly similar sequences are grouped into operational taxonomic units (OTUs), and these OTUs are classified using databases of recognized organisms. OTUs can then be analyzed in terms of presence/absence, abundance, or phylogenetic diversity. To determine biomolecular and metabolic functions present in the community, the total metagenomic DNA may be sequenced and compared with function-oriented databases. Alternatively, sequenced community DNA can be compared to reference genomes. The emerging fields of metatranscriptomics, metaproteomics, and metametabolomics will aid in our understanding of the human microbiome. **Source**: Morgan et al. 2013.

1.4 Colorectal cancer

Tumors of the lower digestive tract, which include colon and rectal cancer, are the third most commonly diagnosed neoplasia with 1.4 million people diagnosed annually (Torre et al. 2015). The World Health Organization estimates an increase of 77% in the number of newly diagnosed colorectal (CRC) cases and an increase of 80% in deaths from CRC by 2030 (Binefa et al. 2014). Sporadic, as opposed to hereditary, CRCs account for approximately 70%-87% of cases (Frank et al. 2017) and genetics can only explain a small proportion of disease incidence (Galvan et al. 2010; Foulkes 2008). This points to the potential role of other variables including lifestyle and environmental factors as disease co-determinants. Reported risk factors associated with CRC include age, tobacco and alcohol consumption, lack of physical activity, increased body weight and diet (Johnson et al. 2013; Wei et al. 2004; Huxley et al. 2009). Of particular importance is the fact that all of these risk factors can directly or indirectly modify the microbiota, and since human colonic cells have high replication rates (10^{11} small intestine epithelial cells being shed each day) (Williams et al. 2015), makes them more susceptible to mutations and consequent carcinogenic transformation, motivating researchers to investigate possible bacterial genera and species associated with CRC.

Due to practical issues related to obtaining the required biopsy samples - from patients and controls - studies have focused their analyses of microbial community compositions on either tissue samples of more specific regions of the lower digestive tract, such as the proximal/distal colon and the rectal tissue, or fecal samples. Fecal- and tissue-associated microbiota have been shown to be significantly different (Durbán et al. 2011; Hong et al. 2011; Mira-Pascual et al. 2015; Flemer et al. 2017) which could lead to a lack of representativeness with respect to the bacterial biofilm of the rectal mucosa (Durbán et al. 2011; Gevers et al. 2014), reflecting the disease state but possibly not the tumor microenvironment. Despite these differences and limitations, fecal microbiota studies have contributed greatly in our understanding of the general gut microbiota composition and its dysbiosis in different scenarios (Wu et al. 2013; Sabino et al. 2016). Whereas, colon and rectal cancers have been routinely studied together as CRC, evidences indicate these to be distinct nosological entities. Differences in embryological origin, anatomy, treatment, metastatic potential, and outcome between colon cancer and rectal cancer have led to discussions as to whether neoplastic lesions of these two anatomical sites should be considered as different diseases, with further dichotomization of colon cancers into distal and proximal (Tamas et al. 2015).

So far, only two studies have investigated the rectal microbiota exclusively (without the inclusion of colon samples). One study (Sanapareddy et al. 2012) investigated rectal biopsy samples from 33 control individuals and 38 individuals with adenomas. Adenoma samples presented increased richness when compared to

controls, as well as increased abundance of the eubacterial genera *Helicobacter*, *Pseudomonas* and *Acinetobacter*. The other study (Araújo-Pérez et al. 2012) used rectal swabs and biopsy samples in order to investigate differences in microbial communities between these two sampling strategies and found significant differences between them.

In work by Zeller et al. 2014, the accuracy of fecal WMS sequencing for CRC detection was compared to the accuracy of the clinically used fecal occult blood test (FOBT). The authors found that both had similar accuracy, however, when both approaches were combined, they observed a 45% increase in sensitivity without losses in specificity (area under the receiver operator curve (AUC) = 0.87; true positive rate = 0.72). Microbial species that contributed with >51% of the total absolute weight in their classification models included two *Fusobacterium nucleatum* subspecies; *Fusobacterium nucleatum vicentii* and *Fusobacterium nucleatum animalis*, and also *Porphyromonas asaccharolytica*, *Peptostreptococcus stomatis*, *Clostridium symbiosum* and *Clostridium hylemonae*, all of whom were enriched in CRC samples. In terms of functional potential, the authors found 24 KEGG modules and 20 CAZy families to significantly differ in abundance in CRC patients, with host cell glycans being significantly more exploited in the CRC-associated microbiome. However, the authors state that using the functional profile alone did not improve classification accuracy in terms of the taxonomic and FOBT classification model (AUC = 0.77). Yu et al. 2015 used a different approach for taxonomic profiling of their samples, creating species-level molecular operational taxonomic units using the mOTU profiling software (Sunagawa et al. 2013) and two other species-level methods; metagenomic linkage groups (MLGs) and mapping reads to the IMG database. They found a significant enrichment of *Eubacterium ventriosum* in healthy controls and of *Parvimonas micra*, *Solobacterium moorei* and *Fusobacterium nucleatum* in CRC samples across all three methodologies. In two of the three methodologies, they found a significant enrichment of *Peptostreptococcus stomatis* in CRC samples. Using random forests for their machine learning approach of CRC detection, the authors obtained AUCs ranging from 0.86-0.96 depending on the classification methodology. For biomarker discovery, the authors used the minimum redundancy maximum relevance (mRMR) feature selection method (Peng et al. 2005) on a catalogue of 140,455 disease-associated genes they created using a Metagenomic Wide Association Study (MGWAS). The authors found 20 genes in their cohort that were strongly associated with CRC status, but when they attempted to use these genes to classify samples of another Danish CRC metagenomic dataset, an AUC of 0.71 was achieved. In work by Feng et al. 2015 (Feng et al. 2015), the authors found 130,715 genes to be differentially abundant between their 3 samples groups (CRCs, advanced adenomas and healthy controls). These genes were then used to create MLGs and were classified by mapping reads to the IMG database using species-level resolution. The authors found a number of MLGs

classified as different *Bacteroides* and *Parabacteroides* species enriched in CRC samples, which also included MLGs classified as *Clostridium symbiosum*, *Bilophila wadsworthia*, *Escherichia coli* and *Alistipes putredinis*. The authors then selected, using random forests and cross-validation, 15 MLGs that performed well on their training set and achieved an AUC of 0.96 on their test set.

Recently, some multi-cohort studies have investigated the link between CRC and the gut microbiome, either through 16S amplicon sequencing or WMS sequencing. Using eleven publicly available 16s rRNA datasets on stool and tissue-based CRC microbiota, Drewes et al. 2017 found that CRC tissues were enriched for invasive biofilms, *Bacteroides fragilis* and oral pathogens which included *Fusobacterium nucleatum*, *Parvimonas micra*, and *Peptostreptococcus stomatis*. In total, 84% of tumors harbored at least one measure of microbial dysbiosis associated with CRC (enrichment of oral microbes, presence of biofilms or enrichment with *B. fragilis*). Using 4 publicly available WMS CRC stool datasets, Dai et al. 2018 identified seven CRC enriched bacteria (*Bacteroides fragilis*, *Fusobacterium nucleatum*, *Porphyromonas asaccharolytica*, *Parvimonas micra*, *Prevotella intermedia*, *Alistipes finegoldii*, and *Thermanaerovibrio acidaminovorans*) that achieved an AUC of 0.8 when used to classify CRC samples and correlated with lipopolysaccharide and energy biosynthetic pathways.

1.5 Gastric cancer

Gastric carcinomas (GC) are the 4th most frequent cause of cancer, accounting for the second highest cancer mortality rate worldwide, with a 5 year survival rate of only 20-30% (Nobili et al. 2011; Cenitagoya et al. 1998; Crew and Neugut 2006). GC mortality has been dropping around the globe, but still remains a serious public health concern in Asia, Andean countries, Eastern Europe and Brazil, where some cities present high prevalence and incidence rates. The main etiologic agents associated with GCs include: low consumption of carotenoids and ascorbic acid, high ingestion of salts and nitrates, infection with *Helicobacter pylori* (HP), infection with the Epstein-Barr virus (EBV), genetic factors, obesity and tobacco use (Crew and Neugut 2006). The chronic gastric infection of *H*. *pylori*, a gram-negative bacilli present in 50% of the world population (Wang and Yang 2013), is nowadays considered the principal etiologic factor of gastric cancer. The prevalence of HP for long periods of time increases gastric cancer susceptibility in up to 75%, due to the activation of different pathways that control the response to oncogenic stimuli in epithelial cells, leading to increased cell proliferation and migration and, most importantly, a constant inflammatory response (Polk and Peek 2010).

Recent studies have investigated the association between microbial communities and GC, despite the knowledge that HP infection alone is a strong driver of gastric carcinogenesis. This is in part due to the hypothesis that other bacteria might play a role in cancer initiation, particularly those that can reduce nitrate. Coker et al. 2017 performed 16S rRNA sequencing of gastric mucosal samples from 81 cases that included superficial gastritis (SG), atrophic gastritis (AG), intestinal metaplasia (IM) and gastric cancer and validated their results using 126 additional mucosal samples. The authors found an enrichment of 21 and a depletion of 10 bacterial taxa in GC compared with SG, which included *Peptostreptococcus stomatis*, *Streptococcus anginosus*, *Parvimonas micra*, *Slackia exigua* and *Dialister pneumosintes* and were correlated with the results from their validation cohort. The authors also reported a significant increase in the number of oral bacteria in GC compared to the other groups and a decrease in species richness when compared to SG. When evaluating the effect of HP on microbial communities, the authors found significantly more interactions among gastric microbes in HP-negative than HP-positive samples, however, no differences were observed in taxonomic diversity or richness in the gastric microbes between HP-positive and HP-negative samples within each disease stage. Ferreira et al. 2018 performed 16S rRNA sequencing of tissue samples (either biopsies or surgical specimens) from 54 patients with gastric carcinoma and 81 patients with chronic gastritis, finding that patients with gastric carcinoma had significantly decreased microbial diversity when compared to patients with chronic gastritis, exhibiting also significant differences in microbial community composition between the two groups. The authors found an enrichment of the genera *Phyllobacterium*, *Achromobacter, Lactobacillus, Clostridium, Citrobacter* and *Rhodococcus* and of the *Xanthomonadaceae* family in gastric carcinoma samples, whereas *Helicobacter*, *Neisseria*, *Prevotella* and *Streptococcus* were enriched in patients with chronic gastritis. The enrichment of *Citrobacter*, *Phyllobacterium*, *Rhodococcus* and *Lactobacillus* in gastric carcinoma patients and enrichment of *Helicobacter* and *Neisseria* in chronic gastritis patients were validated using additional samples analyzed via 16s rRNA amplicon data (79 gastric carcinoma cases) and quantitative polymerase chain reaction (qPCR) of specific genera (15 chronic gastritis patients and 23 gastric carcinoma patients). By calculating a microbial dysbiosis index (MDI), the authors found a higher MDI in gastric carcinoma patients compared to chronic gastritis patients, exhibiting an inverse correlation with alpha diversity and a direct correlation with beta diversity. The authors also found, via functional metagenomic inference, that gastric carcinoma microbial communities had increased nitrate and nitrite reductase functions when compared to chronic gastritis patients. Sung et al. 2016 addressed differences in microbial communities between gastric fluid and gastric mucosa samples from 4 individuals. The authors found that gastric fluid samples possessed higher diversity and richness when compared to paired gastric mucosa samples. The authors also found differences in phyla

composition, with HP and *Proteobacteria* exhibiting higher abundances in mucosa samples compared to gastric fluid samples.

Perioperative chemotherapy associated with surgery is one of the main strategies in the treatment of stage II and III GC patients, which account for the majority of cases. In this context, the addition of some cycles of chemotherapy before and after surgery reduces the risk of tumor relapse and death. However, chemotherapy response is heterogeneous and so far there are no standard methods to assess its efficacy. An increasing amount of evidence shows that the gut microbiota can modulate the host's response to chemotherapeutic drugs, with three main clinical outcomes: facilitation of drug efficacy; abrogation and compromise of anticancer effects; and mediation of toxicity (Alexander et al. 2017). The mechanisms by which the microbiota can modulate the host's response to chemotherapeutic drugs include:

i) Immunomodulation: intestinal microbiota can facilitate chemotherapy-induced immune and inflammatory responses. Routy et al. 2017 and Vétizou et al. 2015 showed the importance of the gut microbiome in the treatment of metastatic melanoma patients and mice models with novel targeted immunotherapies such as anti-PD-L1 and anti-CLTA-4, demonstrating differences in gut microbial profiles in treatment responders and non responders.

ii) Metabolism and enzymatic degradation: direct and indirect bacterial modification of chemotherapeutic drugs might potentiate desirable effects, abrogate efficacy or liberate toxic compounds. Geller et al. 2017 showed that treatment with gemcitabine led to important changes in the gut microbiome of mice, and that gamma-proteobacteria possessing the long isoform of the gene cytidine deaminase were able to rapidly degrade the drug into an inactive form, leading to chemo-resistance.

iii) Reduced diversity: chemotherapy induces changes in the diversity of the mucosal and fecal microbiota through altered biliary excretion and secondary metabolism or associated antibiotic use and dietary modifications (Montassier et al. 2014; van Vliet et al. 2009; Montassier et al. 2015). As a result, pathobionts might predominate, leading to deleterious effects such as diarrhoea (van Vliet et al. 2009).

1.6 Motivation and aims

The gut microbiome can have direct and/or indirect effects in cancer initiation, promotion and response to therapy (Vétizou et al. 2015; Geller et al. 2017; Rubinstein et al. 2013; Kostic et al. 2013). There are different proposed mechanisms as to how these effects may occur: through dysbiosis, production of bacterial toxins, secondary bacterial metabolites, among others. Motivated by these broad and important microbial effects, in this thesis we studied the association between the gastric and the gut microbiome and cancer from different perspectives, investigating whether alterations to microbial profiles were associated with disease status or an adverse response to treatment.

1. Due to the absence of studies that evaluated the rectal tissue microbiota in the context of rectal cancer exclusively, in the second chapter of this thesis, entitled "*Tissue-associated 16S rRNA community profiling of rectal carcinoma patients*", we used biopsy tissue samples to profile bacterial communities that were associated with the rectal mucosa through 16S rRNA amplicon sequencing, characterizing the bacterial diversity, community structure and membership when contrasting healthy and diseased states.

2. Recent multi-cohort studies have investigated the link between the microbiome and CRC but had limitations that included low sample size, use of the low-resolution 16S rRNA amplicon sequencing approach, or low diversity of the targeted populations. Therefore, in the third chapter of this thesis, entitled "*Combined metagenomic analysis of colorectal cancer datasets*", we identified reproducible CRC microbial biomarkers and assessed prediction accuracies for CRC detection across populations, datasets, and conditions through fecal WMS sequencing.

3. The gut microbiota can modulate the host's response to chemotherapeutic drugs, with studies showing an association between gut microbial composition and both immunotherapy response and chemotherapy efficacy. Therefore, in the fourth chapter of this thesis, entitled "*Neoadjuvant chemotherapy treatment in gastric cancer patients reveals shifts in gastric microbial communities*", our aim was to assess whether gastric microbial communities were modulated by neoadjuvant chemotherapy treatment, aiding in the discovery of microbial markers of response to treatment that could help improve treatment decisions and outcomes in patients with a potentially curable disease.

Chapter 2. Tissue-Associated 16S rRNA Community Profiling of Rectal Carcinoma Patients

In face of the microbiota gradient found in the human digestive tract (Flemer et al. 2017; Zhang et al. 2014; Gao et al. 2015), the possibility that tissue-associated microorganisms could play a more direct role in cancer initiation and development and the absence of studies investigating the microbiome of rectal tumors exclusively, this chapter focuses on investigating microbial communities present in rectal tissue samples, when contrasting healthy and tumor samples, through 16S rRNA amplicon sequencing. The results presented in this chapter are a modified version of the results published by Thomas et al. 2016.

2.1 Materials and Methods

2.1.1 Cohort

With the help of Dr. Eliane Camargo Jesus, a total of 36 subjects were included after approval by AC Camargo Cancer Center's ethics review board (ACCCC - 1614/11, January 30th, 2012). Tissue biopsies were collected from subjects belonging to one of the following groups:

Non-cancer subjects – (Non-Cancer, NC, n= 18): All subjects had medical indication of exploratory colonoscopy due to complaints, such as bleeding, abdominal pain, constipation and chronic diarrhea. No subjects had personal or familial history of colorectal cancer or colitis (either ulcerative, Crohn's, radiation or infectious colitis, chronic inflammatory illnesses), previous colonic or small bowel resection, nor previous colon adenomas or familial polyposis syndrome. Only individuals with complete colonoscopies that allowed the full visualization of the entire colon and showed no significant clinical alterations were included.

Colonoscopy and biopsy procedures for the NC subjects - All patients received standard instructions for preparation for colonoscopy that included consumption of 500 ml of mannitol for bowel cleansing, luftal and bisacodyl. Eligible subjects gave written informed consent to provide colorectal biopsies, had their anthropometric measures taken and answered questions about diet, consumption of alcohol and tobacco. Colonoscopy was performed using a Pentax videoscope model FC38LX. During biopsy procurement, the rectum was inflated with air and care was taken not to use any suction during advancement of the scope to 7-8 cm from the anal verge. Sterile biopsy forceps were not taken out of the channel of the scope until an area that was completely clear of stool was seen with clear pink mucosa. Biopsies were taken with 2.2 mm sterile standard forceps.

Patients diagnosed with Rectal Adenocarcinomas (Rectal-Cancer, RC, n= 18): Tumor specimens, located in the higher (N=15), mid (N=22) and lower rectum (N=1), were obtained from surgeries to remove the tumor mass. All subjects belonging to this group were recruited with the help of Drs. Samuel Aguiar Junior and Ademar Lopes at AC Camargo Cancer Center's Pelvic Surgery Department, in São Paulo, Brazil. We included patients that were diagnosed with rectal adenocarcinoma [tumors of stage pT1 or pT2 low- or mid-straight, pT1 or pT2 or pT3 high-straight], that had not undergone any neoadjuvant therapy and had their tumors surgically resected at the Pelvic Surgery Department, AC Camargo Cancer Center, with diagnosis confirmed by Dr. Maria Dirlei Begnami of the Pathology Department of the same institution. After the histopathologic confirmation of rectal adenocarcinoma diagnosis, surplus samples were macrodissected and used for DNA extraction and bacterial community profiling. *Exclusion criteria* were: patients subjected to neoadjuvant therapy prior to tissue collection; patients reporting inflammatory bowel diseases or with hereditary cancer syndromes. We also excluded all subjects (cases and controls) who reported the use of antibiotics for at least 4 weeks prior to sample-collection.

2.1.2 DNA extraction

DNA extraction started after incubating the samples for 18 hours in 600µl of a lysis buffer (Qiagen) and 15µl of proteinase K (20µg/µl) at 55ºC. After this period, DNA samples were extracted using a standard phenol chloroform protocol, followed by ethanol precipitation, quantification using a spectrophotometer (Nanodrop – Thermo Scientific) and visualized on 2% agarose gels to inspect DNA integrity.

2.1.3 PCR amplification and sequencing of the V4-V5 region of 16S rRNA gene

The V4-V5 region was amplified using a primer set designed to generate amplicons compatible with the chemistry available for the Ion Torrent PGM platform, that allowed ~400 nt of high quality sequences (Ion PGM Sequencing 400 Kit). Coverage of the primer set was evaluated using the Ribosomal Database Project's (RDP - Release 11.2), ProbeMatch (Cole et al. 2014) and the ARB Silva's (Release 115) TestPrime (Klindworth et al. 2013). The forward primer (5'-AYTGGGYDTAAAGNG-3') and reverse primer (5'-CCGTCAATTCNTTTRAGTTT-3') corresponded to positions 562 and 906, respectively, of the *Escherichia coli* 16S rRNA gene.

Three 50µl amplification replicate reactions were performed per sample, each containing: 2.5µM of each primer; 25µl of Kapa Hotstart High Fidelity Master Mix (Kapa Technologies) and 25ng of genomic DNA (gDNA). Thermocycling conditions were: 95ºC, 3 min; 98ºC, 15 sec and 40ºC, 30 sec for 35 cycles; followed by a last extension step at 72ºC for 5 min. Amplicons of the three reactions from each subject

were pooled and purified using a MinElute PCR Purification Kit (Qiagen). The purified products were run on 1.5% agarose gels and gel bands within the expected amplicon range were excised using sterile and disposable scalpels and purified using the Qiaquick gel extraction kit (Qiagen) to remove artifacts, primer-dimers and non-specific bands. Amplicons were end-repaired and Ion Torrent adaptors with barcodes were ligated. Equimolar amounts of amplicons from each sample were pooled, using the Ion Torrent qPCR quantitation kit (Thermo Scientific), and used for emulsion PCR. All samples were sequenced on the Ion torrent PGM platform (Thermo Scientific) using two 318 v2 chips. Samples from both groups were processed simultaneously, to avoid possible batch effects.

2.1.4 Sequence analysis

Sequences processed by the Ion Torrent server (v3.6.2) were used as input into the *Qiime* (*Quantitative insights into microbial ecology*) software package (Version 1.6.0) (Caporaso et al. 2010). We first removed sequences with an average quality score <20 using a 50nt sliding window. Then, we identified barcodes used for subject-assignment, allowing a maximum of 2 mismatches, and discarded sequences with no barcodes, and <200nt or >500nt after barcode removal. PCR primers identified at the start or at the end of the reads, allowing a maximum of 4nt mismatches, were trimmed and sequences with no identifiable primers were discarded. After primer trimming we removed all sequences below 200nt and the remaining sequences were used as input for downstream analysis.

Filtered sequences were clustered with 97% identity using UPARSE (implemented in *USEARCH* v7) (Edgar 2013) and the seed sequence of each cluster was picked as a representative. Chimeric sequences (and clusters) were identified using UCHIME (Edgar et al. 2011) and the Broad Institute's chimera slayer database (version microbiomeutil-r20110519) and excluded from further analysis. The RDP classifier (Wang et al. 2007), as implemented within the *Qiime* interface (default parameters), was used to assign taxonomic ranks using a minimum confidence value of 80% and, subsequently, to each operational taxonomic unit (OTU). Unless otherwise stated, OTUs that occurred in less than 25% of all samples and with less than 3 reads were not considered.

In this chapter as well as in subsequent chapters we use the following definition for relative abundance of a taxon: for a given sample, it is the number of sequenced reads associated with that taxon divided by the total number of reads that have passed quality control, multiplied by 100.

2.1.5 Alpha and beta diversity analysis

We rarefied the OTU table to 17,414 sequences per sample in order to calculate species diversity, using the Shannon-Weaver index (Shannon 1948) and

the Simpson index (Simpson 1949), and richness (by using the observed species) implemented in the R Phyloseq package (McMurdie and Holmes 2013).

For beta diversity analysis, OTU-representative sequences were aligned using PyNAST (Caporaso, Bittinger, et al. 2010) against the aligned *greengenes* core set (DeSantis et al. 2006) with *Qiime* default parameters, and the alignments were lanemask-filtered (Lane 1991). A phylogenetic tree was built using FastTree (Price et al. 2009), weighted and unweighted UniFrac (Lozupone and Knight 2005) distances were calculated and a distance matrix was generated. Using the R phyloseq package, distance matrices were used to calculate coordinates for principal coordinate analysis (PCoA).

Enterotypes. Community types of each sample were analyzed by the Dirichlet multinomial mixture model-based method (Holmes et al. 2012) using rarefied genera level counts of 16S rRNA sequencing reads. Partitioning around medoids (PAM) *enterotyping* was performed in R using genera level relative abundances and the "cluster" package (Maechler et al. 2018). We applied 4 distance metrics: Weighted UniFrac, Unweighted UniFrac, root Jensen-Shannon divergence and Bray-Curtis and assessed the quality of the clusters using prediction strength (Tibshirani and Walther 2005), silhouette index and the Caliński-Harabasz statistic (Rousseeuw 1987) using the "fpc" R package.

2.1.6 Differential abundance analysis

To investigate differences in OTU, phyla and genera abundances between both groups, raw counts were normalized then log transformed using the normalization method below, as performed by a previous study (Sanapareddy et al. 2012):

$$Normalized\ count = log10(\ (\ raw\ count \div number\ of\ sequences\ in\ that\ sample)\ \times average\ number\ of\ sequences/sample\ +\ 1\ )$$

We also evaluated high-level phenotypical differences in microbial composition between both groups. Quality control passed sequences were closed-reference picked at 97% identity using UCLUST_Ref (Edgar 2010) and the green genes core set (Version 13.5). The resulting OTU table was rarefied to 13,944 sequences and submitted to BugBase (Ward et al. 2017) in order to calculate differences between both groups in terms of microbial phenotypes.

2.1.7 Data validation

With help of our collaborators Ana Maria Camargo Aranha and Paola Avelar Carpinetti, we detected and quantified the absolute number of 16S rRNA *B. fragilis* copies in our samples using the *QX200™ Droplet Digital™ PCR System* (Bio-Rad). The primers used to amplify the *B. fragilis* 16S rRNA gene were: BF-fwd

5’-TCRGGAAGAAAGCTTGCT-3’ and BF-rev 5’-CATCCTTTACCGGAATCCT-3’ (Tong et al. 2011) and to ensure further specificity, a labeled probe BF-p 5’[FAM]-ACACGTATCCAACCTGCCCTTTACTCG-3’ [BHQ1] (Tong et al. 2011) was included in the reaction. We used a commercial RNAseP *Copy Number Reference Assay* (Thermo-Fisher Scientific) to detect and quantify human DNA. Microdroplets (~20.000/reaction) were generated on the Bio-Rad *QX-100* following the manufacturer's instructions. RNAse P and *B. fragilis* ddPCR were performed in 96 well-plates, in a final volume of 20µl, containing: 15ng of total DNA, 10ul of ddPCR supermix for probes (Bio-Rad), 8 pmol of each PCR BF-primer and 2 pmol of the BF-probe, or 1µl of RNAse P assay. PCR conditions were: 50ºC- 2 minutes; 95ºC- 10 minutes; 95ºC- 15 seconds and 60ºC- 1 minute for 40 cycles. After cycling, the 96-well plate was immediately transferred on a *QX200 Droplet Reader* (Bio-Rad), where flow cytometric analysis determined the fraction of PCR-positive droplets versus the number of PCR-negative droplets in the original sample. Data acquisition and quantification was carried out using *QuantaSoft Software* (Bio-Rad). To ensure the accuracy of the results, a minimum of 10,000 acceptable droplets per reaction were required for quantification using the *QuantaSoft software*. Samples yielding a minimum of 3 positive droplets from 10–15,000 droplets analyzed were scored as positive.

2.1.8 Immunohistochemistry

Immunohistochemistry was performed by our collaborator Rafael Malagoli Rocha in an automated Benchmark platform (Ventana Medical Systems) for Anti-*B. fragilis* LPS antibody (mouse monoclonal - Abcam 1265/30) in whole slide tissues. Alkaline phosphatase conjugated to secondary polymeric system was used for IHC visualization. The selection of positive and negative samples was guided by the high-throughput sequencing (HTS) data and used to confirm the presence of *B. fragilis* in the sample set. The primary antibody was omitted to evaluate background staining.

2.1.9 Statistical analysis

Wilcoxon tests were used to compare mean differences between tumor and biopsy samples for phyla, genera and OTU log-abundances. *P-values* were corrected for multiple testing using the Benjamini and Hochberg procedure (Benjamini and Hochberg 1995). Fold changes for each genera/OTU were calculated using:

$$Log2FC = log2(RC\ average + 1) - log2(NC\ average + 1)$$

Chi-Square tests were performed on subject's categorical data such as gender, alcohol and tobacco use and vital status. Student t-tests were performed to compare differences in the means between both groups for age, height, weight, BMI and alpha diversity. We used ANOSIM and ADONIS (Oksanen et al. 2018) to compare differences in beta-diversity between groups using 3 distance metrics weighted UniFrac, unweighted UniFrac and Bray-Curtis for categorical and numerical variables, respectively. Linear models were built using normalized counts at the genera and OTU level to investigate associations with clinical-pathological characteristics of rectal-cancer samples, such as lymph node and perineural neoplastic invasion status. Unless otherwise stated, values were reported as mean ± standard deviation and *p-values* <0.05 were considered statistically significant. All calculations were performed within the R statistical computing environment (R Core Team 2018) unless otherwise stated.

2.2 Results

We analyzed tissue-associated bacteria from mucosal biopsies of 18 non-cancer controls and 18 rectal adenocarcinoma tumors using 16S rRNA high throughput amplicon sequencing. With the help of Dr. Helano Carioca Freitas, we found no significant differences between rectal-cancer and non-cancer subjects regarding age and gender distribution, tobacco and alcohol use and other risk factors (Table 1). All samples consisted of rectal-biopsies. The biopsies of individuals with no tumor lesions derived from the mid rectum and were distributed along the ~12cm-long human rectum, with most samples deriving from the higher-mid rectum (94%). Effect size analysis using pairwise distances (Kelly et al. 2015) between both groups revealed an ω^2 ranging from 0.13 – 0.26, depending on the metric, with PERMANOVA *p-values* <0.001, indicating that this sample size allows the observation of significant microbial differences between the two sample groups.

Our analyses indicated that the PCR primers used here (V4-V5 region of the 16S rRNA gene) covered 84.4% and 52.1% of all eubacterial sequences present in the ARB SILVA database and the Ribosomal Database Project, respectively (**Appendix 1**). Coverage rates were evenly distributed among most bacterial phyla, except for *Verrucomicrobia*, where coverage rates were 21% and 10.9%, dropping below the 75% and 48% averages of taxa present in the SILVA and RDP databases, respectively.

2.2.1 Sequence analysis

A total of 12,078,140 sequence reads were generated, with a mean sequence length of 304.5nt ± 97.34nt (standard deviation - std). After quality filtering and primer trimming, 5,593,020 (46.3%) sequences remained, with an average of 155,361 sequences/sample and a mean sequence length of 315nt ± 30nt.

Table 1 – Subject and sample data

Demographic	Non-cancer (n=18)	Rectal-cancer (n=18)	*p-value*
Age	55.2 ± 15.7	59.3 ± 8.8	0.348
Gender (%) Female Male	 9 (50) 9 (50)	 8 (44) 10 (56)	1
Height	1.65 ± 0.08	1.70 ± 0.09	0.1
Weight	73 ± 14.1	73.8 ± 13.5	0.87
BMI	26.6 ± 3.7	25.3 ± 3.6	0.29
Alcohol Use (%) Yes No Undetermined	 8 (44) 10 (56) 0 (0)	 5 (28) 12 (67) 1 (5)	0.568
Tobacco Use (%) Yes No Undetermined	 12 (67) 6 (33) 0 (0)	 6 (28) 11 (62) 1 (5)	0.129
Pathological tumor size staging (%) pT2 pT3	N.A.	 5 (28) 13 (72)	N.A.
Pathological lymph node metastasis staging (%) pN0 pN1 pN2	N.A.	 11 (62) 3 (16) 4 (22)	N.A.
Distant metastasis staging (%) M0	N.A.	18 (100)	N.A.
Invasion (%) Perineural Angiolymphatic	N.A.	4 (22) 14 (78)	N.A.
Vital Status (%) Alive Deceased	18 (100) 0 (0)	17 (95) 1 (5)	1

N.A. – Not applicable

When all individuals were considered, a total of 3,222 OTUs were obtained. Thirty-one (0.7%) OTUs were identified as chimeras by UCHIME and 209 (4.7%)

could not be assigned to a taxonomic rank. After filtering OTUs with less than three sequences and not present in at least 25% of all samples (NC and RC combined), 1,327 OTUs remained.

2.2.2 Alpha and beta diversity

We observed significantly higher species richness and species diversity in rectal cancer samples compared to controls. This was observed for the number of OTUs, the Shannon index and the Simpson Index (*p-values* = 0.002, <0.001 and <0.001, respectively) (**Figure 3A-B**). When we stratified rectal-cancer samples into smaller (pT2) and larger tumors (pT3), we observed an increase in species richness, with an average of 280 and 366 OTUs, respectively, compared to 236 OTUs in NC; however this effect reached no statistical significance between pT2 and pT3, maybe because of the reduced number of pT2 samples (N=5, compared to N=13 for pT3) (**Figure 3B**).

Using three distance metrics we observed consistent and statistically significant differences between the sample groups when considering cancer status (Bray-Curtis, Unweighted and Weighted UniFrac; *p-value* 0.001; ANOSIM using 999 permutations), but not for any other categorical or numerical variable, which included amplicon library construction, age, gender, BMI, alcohol and tobacco use (**Figure 3D; Appendix 2**).

Enterotyping analysis using a Dirichlet multinomial mixture model divided our cohort in two clusters (**Figure 4A-C**). Enterotype I was significantly enriched for rectal-cancer samples, whilst enterotype II was composed mostly of non-cancer samples (*p-value* 0.0001, Fisher's exact test). Enterotype I had higher abundances of *Bacteroides, Clostridiales*, *Dorea* and other genera, whilst enterotype II was characterized by elevated amounts of *Pseudomonas* and *Brevundimonas* (**Figure 4D**). When using the PAM based enterotyping method and criterion adopted by a meta-analysis of human enterotypes (Koren et al. 2013), we found two enterotypes with prediction strength above 0.9 (meaning that 90% of the data points fall within the cluster and 10% are outliers) using the Weighted UniFrac distance (**Appendix 3**).

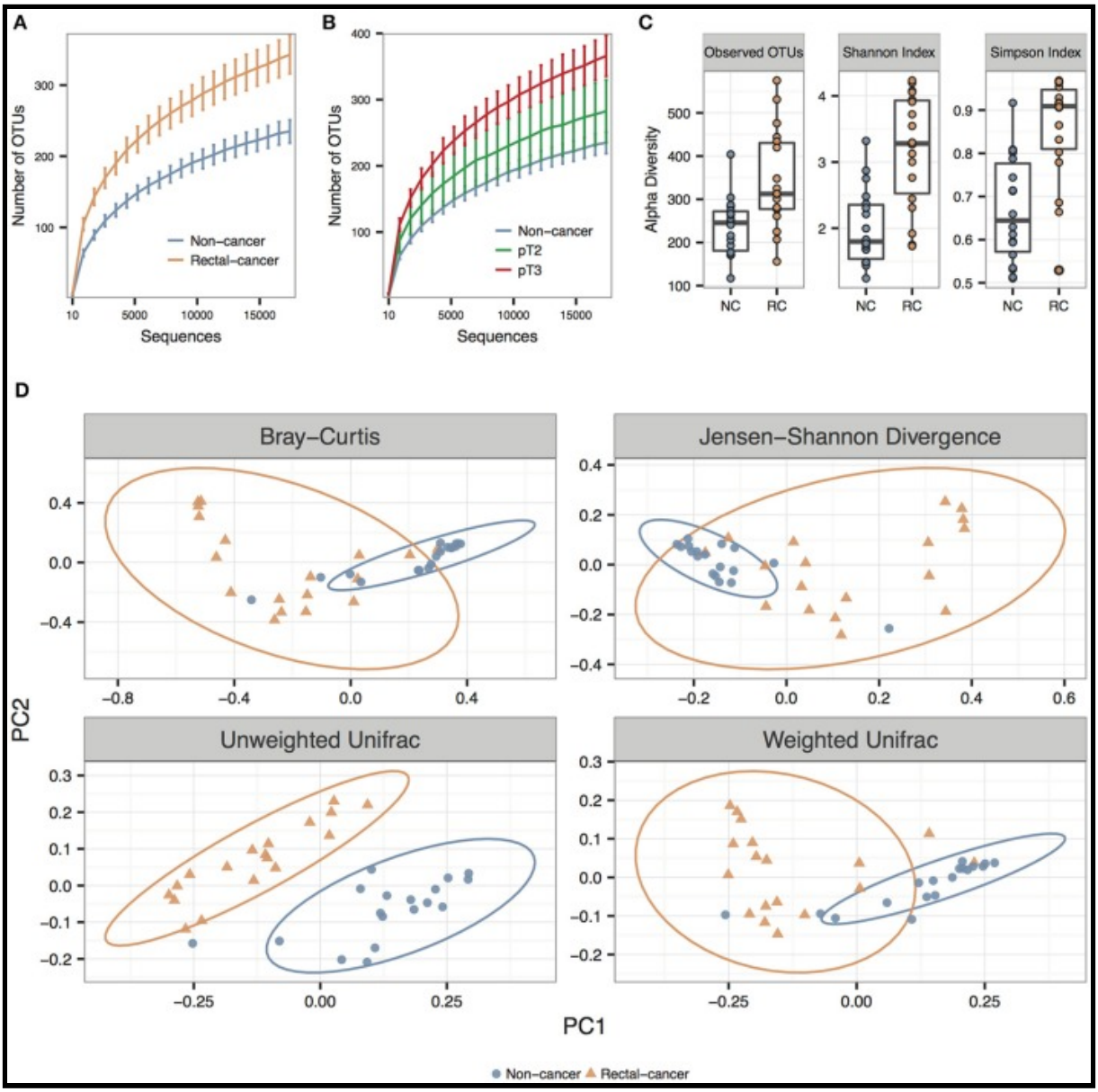


Figure 3. **Alpha and beta diversity for non-cancer and rectal-cancer samples. (a)** Rarefaction curves showing the average number of observed OTUs for both groups. Error bars represent ± standard error of the mean. Blue: non-cancer samples; red: rectal-cancer samples. **(b)** Rarefaction curves showing the average number of observed OTUs for NC samples and for smaller (pT2) and larger rectal tumors (pT3). Error bars represent ± standard error of the mean. Blue: non-cancer samples; red: rectal-cancer samples. **(c)** Boxplots showing alpha diversity in rectal-cancer samples and non-cancer samples using different metrics (Observed OTUs, Shannon index and Simpson index). **(d)** Principal Coordinate Analysis (PCoA) ordination plots for four distance metrics (Bray-Curtis, Jensen-Shannon Divergence, Weighted and Unweighted UniFrac). Ellipses represent the 95% confidence level assuming a multivariate t-distribution. **Source:** Thomas et al. 2016**.**

2.2.3 Global signatures of the microbial community

We observed a significant difference in the log abundances of 6 out of 12 detected phyla between both groups (**Appendix 4**). The most abundant phyla identified were (in decreasing order) *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Fusobacteria*, *Actinobacteria,* and *Verrucomicrobia*. In non-cancer samples, we observed higher log abundances of *Actinobacteria*, *Cyanobacteria, Proteobacteria*

and *Planctomycetes,* whose presence was detected in 9/18 NC samples, with an average log abundance of 0.54 and was absent from all RC individuals (*p-value* < 0.001). In rectal-cancer we found greater log abundances of *Bacteroidetes* and of the much less known candidate phylum *OD1* (also known as *Parcubacteria)*, whose presence was detected in 14/18 RC samples with an average log abundance of 0.71 versus 1/19 NC samples and an average log abundance of 0.02 (*p-value* < 0.001).

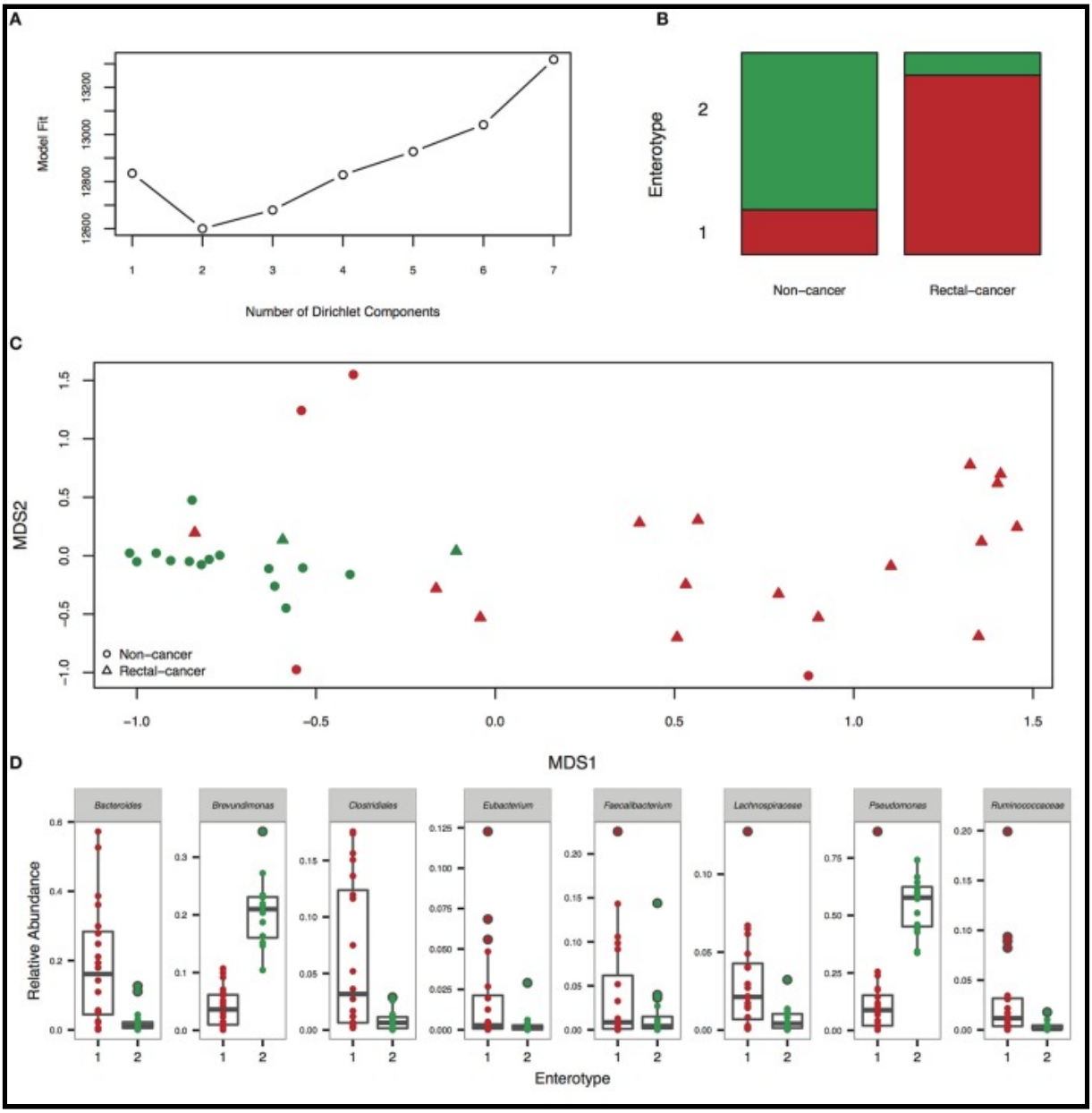


Figure 4. Enterotyping analysis reveals the presence of two community types. (a) Fitting to the Dirichlet Multinomial Mixture model indicates optimal classification into two community types. **(b)** Distribution of rectal-cancer samples and non-cancer samples in both enterotypes (p = 0.0001, Fisher's exact test). **(c)** Non-metric dimensional scaling (NMDS) ordination plot of Jensen–Shannon divergence values between samples. Red, community type-1; green, community type-2. **(d)** Relative abundances of the top 8 most divergent genera between the two community types. **Source:** Thomas et al. 2016.

Genera log abundances. At the genus level, 86 out of 260 genera (33%) showed significant differential log abundances between both groups (**Figure 5A**). The top five genera with differential log abundances between the groups were *Bacteroides*, *Phascolarctobacterium*, *Odoribacter*, *Parabacteroides, Desulfobrio* (more abundant in the cancer group) and *Lactobacillus, Pseudomonas*, *Bacillus, Escherichia*, *Acinetobacter* (more abundant in the non-cancer set) (**Figure 5B**).

Of the 1,492 OTUs identified, 163 (10.9%) were found to have significant differential log abundances between both groups (**Figure 5C**). Three OTUs assigned to the genus *Bacteroides,* two belonging to *B. fragilis* and one to *B. uniformis,* as well as OTUs assigned to *Bilophila* sp. and *Fusobacterium* sp., were significantly more abundant in rectal-cancer samples (**Figure 5D**). In non-cancer samples, OTUs assigned to *Alcaligenes faecalis, Bacillus cereus*, *Lactobacillus delbruecki*, *Prevotella melaninogenica* and *Pseudomonas ssp* had higher log abundances compared to rectal-cancer samples. Four OTUs belonging to the *Bacilli* class were more abundant among non-cancer samples, including *Lactobacillus delbrueckii* (**Figure 5D**).

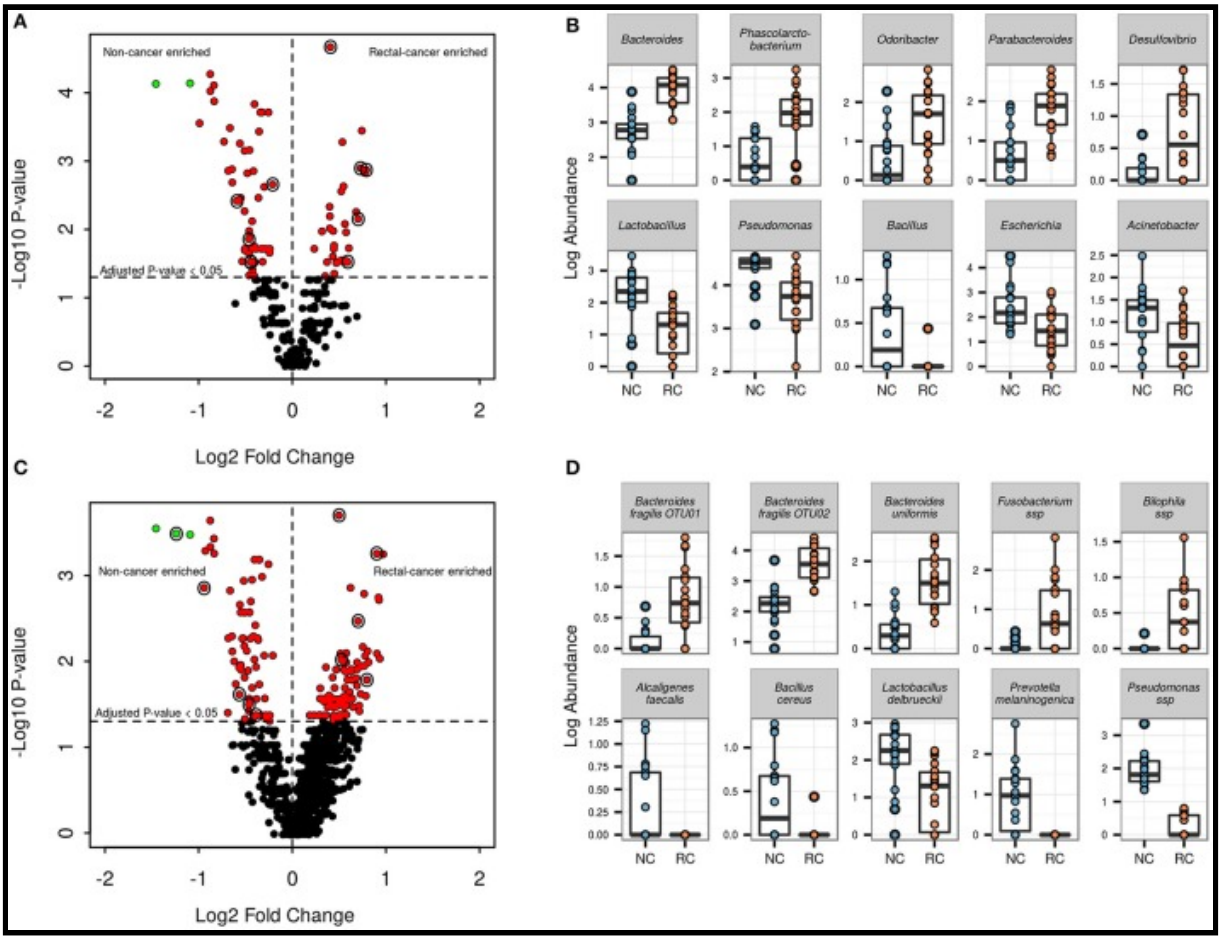


Figure 5. Genera and OTU level differential abundance signatures. (a) Volcano plot for all 260 genera found in our samples. Red points indicate genera with an adjusted *p-value* <0.05; green points indicate genera with an adjusted *p-value* <0.05 and log2FC >1. Points circled in black are genera shown in the adjacent boxplot. **(b)** Boxplots showing log abundances for 5 genera with significant increases (top) and 5 genera with significant decreases in rectal-cancer samples (bottom). **(c)** Volcano plot for 1492 OTUs found in our samples. Points color scheme is the same as in **(a)**. **(d)** Boxplots showing log abundances for 5 OTUs with significant increases (top) and 5 OTUs with significant decreases in rectal-cancer samples (bottom). **Source:** Thomas et al. 2016.

2.2.4 Digital droplet polymerase chain reaction of *Bacteroides fragilis* abundance

As two OTUs classified as *Bacteroides fragilis* were among the smallest *p-values* found and with the highest fold change between the groups, we designed a specific ddPCR assay for *B. fragilis* in order to verify the validity of the results using an alternative approach. As can be seen in **Figures 6A** and **6B**, we observed the expected correlation (R^2 = 0.78) between both methods and confirmed the higher ratio of *B. fragilis*/human DNA in rectal cancer samples, validating the results of our sequencing approach (*p-value* = 0.04, Wilcoxon Rank-Sum Test). To further evidence the presence of *B. fragilis* in tumor specimens, we performed an immunohistochemistry assay on 3 rectal-cancer samples using an anti-*B. fragilis* LPS antibody and found that this bacterium was present in rectal-cancer tissue (Figure **6C-D**).

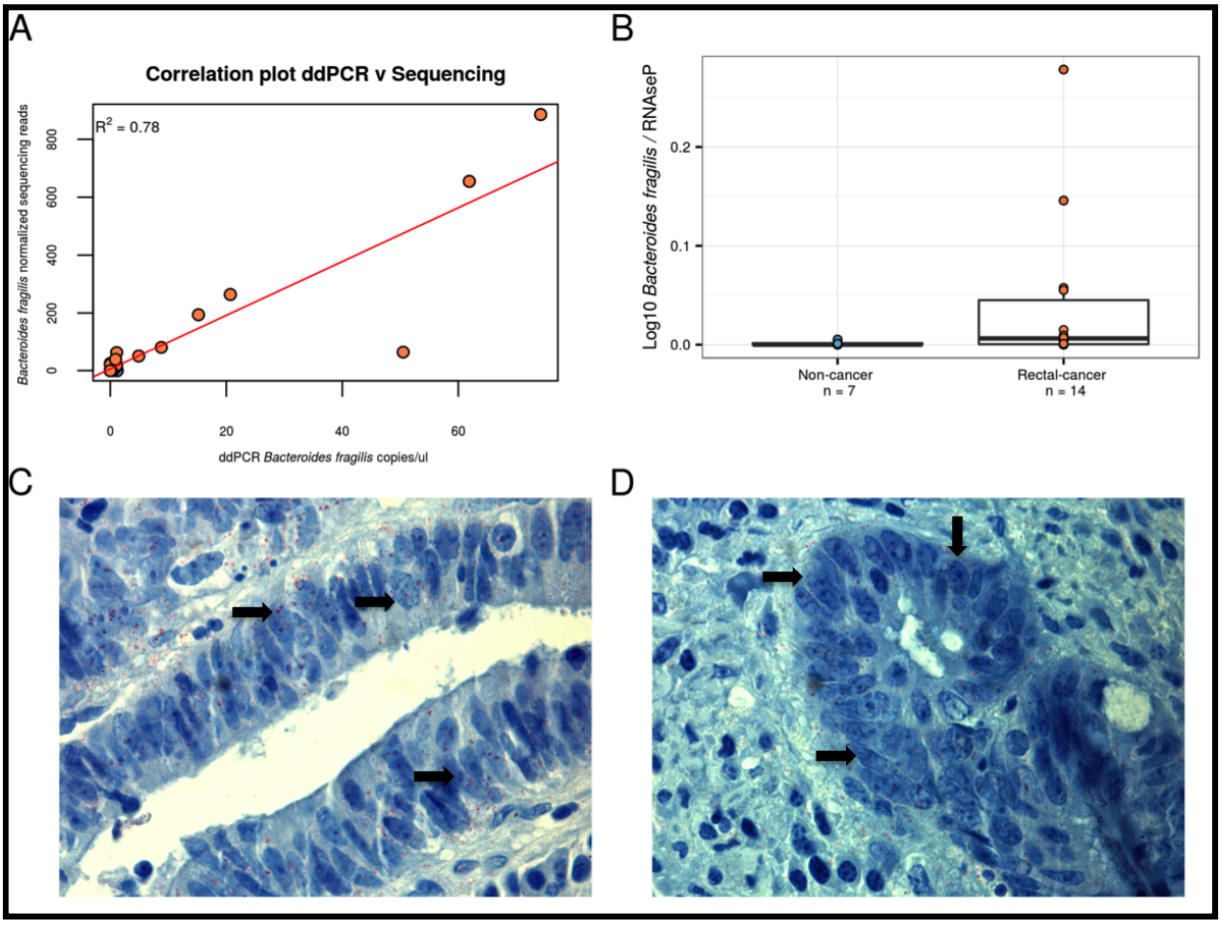


Figure 6. Alternative approaches demonstrating the presence of *B. fragilis*. **(a)** *B. fragilis* ddPCR quantification correlates with HTS-derived data. Using linear regression we obtained a correlation of R^2 = 0.78, $p < 0.001$. Blue: non-cancer samples; red: rectal-cancer samples. **(b)** Boxplot showing log10 of the ddPCR ratio found for *B. fragilis* after normalizing for RNAseP values for both groups. Blue: non-cancer samples; red: rectal-cancer samples. **(c,d)** Immunohistochemisty analysis of two *B. fragilis*-positive rectal-cancer samples, demonstrating the presence of this microbe (antibodies are labeled in red and shown with arrows) using magnification of 1000X. **Source:** Thomas et al. 2016.

2.3 Discussion

We observed increased species-diversity and -richness among rectal-cancer samples. Higher species-diversity and -richness were seen in rectal tissue samples from adenomas compared to normal samples (Sanapareddy et al. 2012) and CRCs versus adenomas (Nakatsu et al. 2015) and increased richness was found in CRCs compared to both adenomas and controls (Mira-Pascual et al. 2015). However, when looking at fecal samples, studies have had conflicting results. One study found increased richness of both genes and genera along the adenoma-carcinoma transition (Feng et al. 2015), whereas another found a decrease in diversity when comparing carcinoma samples and normal controls (Ahn et al. 2013) and a third found no differences between controls, adenomas and carcinomas (Zeller et al. 2014). It is noteworthy to state that these fecal studies grouped proximal and distal colon cancers together with rectal cancers, which could have led to differences in their results. We should also note that the five cases of early-stage lesions (pT2) showed, on average, intermediate microbial richness, when compared to non-cancer biopsies and a more advanced neoplastic stage (pT3). This suggests that increased species richness of cancer lesions could have an early role in rectal carcinogenesis.

Alterations found at the phylum level include higher levels of *Cyanobacteria* (possibly *Melainabacteria*) (Soo et al. 2014), *Actinobacteria*, *Bacteroidetes*, *OD1*, *Proteobacteria* and *Planctomycetes* in the RC-group. We should note an important abundance difference for bacteria of the candidate phylum *OD1* (*Parcubacteria*). These highly adapted organisms have not been isolated in vitro yet; they have small genomes (<1Mb) and reduced metabolic properties identified in a range of anoxic environments. The absence of biosynthetic capabilities and DNA repair enzymes, derived from the genomic analyses of some *OD1* bacteria, suggests a role as ectosymbionts (Nelson and Stegen 2015). However, the putative role of these microbes in rectal cancer remains to be determined. A second phylum, *Planctomycetes*, which are atypical bacteria relatively close to *Verrucomicrobia* and more frequently observed in aquatic environments (such as saltwater, fresh water and acidic mud) (Fuerst and Sagulenko 2011), also showed potential as a biomarker for RC, with striking differences between the groups.

Bacteroides fragilis, a symbiotic organism common to the human intestinal tract, was found to be more abundant in rectal-cancer samples seen by 16S rRNA sequencing and confirmed by ddPCR. Other studies that investigated tissue-associated bacteria also found increased abundance of *B. fragilis* in tumor samples (Zeller et al. 2014; Nakatsu et al. 2015; Drewes et al. 2017). *B. fragilis* has been identified as an important human intestinal symbiont and has been suggested to act as a “keystone pathogen” in the development of CRC (Hajishengallis et al. Curtis 2012). *B. fragilis* is an obligate anaerobe and is a minority member of the normal colonic microbiota with a propensity for mucosal adherence (Sears et al.

2014). Previous reports have linked enterotoxigenic *B. fragilis* (ETBF) to human diarrheal illnesses and increased tumorigenesis in an IL-23-dependent and STAT3-dependent manner (Wick et al. 2014). The toxin fragylisin, produced by ETBF, is a zinc-dependent metalloprotease that triggers NF-kB signaling and cleaves E-cadherin, and has been suggested to be oncogenic (Wu et al. 2009). Bacterial genera known for their role in butyrate production, such as *Ruminococcus*, *Roseburia,* and *Butyricimonas* were more abundant among rectal-cancers, differing from results reported so far (Bultman and Jobin 2014). An OTU assigned to *Bilophila,* a bile-resistant, strictly anaerobic bacterial genus, was also more abundant among rectal-cancer samples, and evidence suggests that products of bacterial bile acid conjugation, secondary bile acids, are carcinogenic (McGarr et al. 2005; Ridlon et al. 2014). *Desulfovibrio*, a commensal sulfate-reducing bacterium, may contribute to mucosal inflammation through hydrogen sulfide production, a resulting by-product of sulfated mucin metabolism (Earley et al. 2015). *Phascolarctobacterium,* known to produce propionate via succinate fermentation, was also increased among cancer samples. On the other hand, we found that *Lactobacillus delbrueckii* was more abundant in non-cancer samples. Probiotic *Lactobacilli* can modify the enteric flora and are thought to have a beneficial effect on enterocolitis. Treatment of IL-10-deficient mice with the probiotic *Lactobacillus salivarius ssp*. reduced the intensity of mucosal inflammation and the incidence of colon cancer from 50% to 10% (O'Mahony et al. 2001).

Chapter 3. Combined metagenomic analysis of colorectal cancer datasets

Recent multi-cohort studies have investigated the link between the microbiome and CRC (Dai et al. 2018; Drewes et al. 2017) but have suffered from low sample size and technical limitations such as taxonomic resolution and cross-dataset comparability. The recent availability of multiple whole-metagenome shotgun datasets of CRC cohorts (Zeller et al. 2014; Feng et al. 2015; Vogtmann et al. 2016; Hannigan et al. 2017; Yu et al. 2017) presents a distinct opportunity for studying the cancer-associated microbiome. This chapter describes the methods and results obtained by combining multiple small whole-metagenome shotgun CRC cohorts of potentially low generalizability, to obtain a better representation of the spectrum of cancer cases and controls, providing strain-level and functional potential resolution, aiding in the understanding of the microbiome's possible role in cancer initiation and promotion. The results presented in this chapter are a modified version of the results by Thomas, Manghi et al. 2019.

3.1 Materials and methods

3.1.1 Cohorts

The study was approved by the local ethics committees (Cohort1: Ethics committee of Azienda Ospedaliera "SS. Antonio e Biagio e C. Arrigo" of Alessandria, Italy, protocol N. Colorectal_miRNA_CEC2014 and Cohort2: Ethics committee of European Institute of Oncology of Milan, Italy, protocol N. R107/14-IEO 118) and informed consent was obtained from all participants.

For Cohort1, samples were collected from patients recruited in a hospital-based manner at Clinica S. Rita in Vercelli, Italy with the help of our collaborators Drs. Alessio Naccarati, Sonia Tarralo, Barabara Pardini, Antonio Francavilla, Gaetano Gallo and Mario Trompetto. Patients with hereditary CRC syndromes or with previous history of CRC, with uncompleted or poorly cleaned colonoscopy were excluded from the study. Patients were recruited at initial diagnosis and had not received any treatment prior to fecal sample collection and subjects reporting the use of antibiotics were excluded from the study. On the basis of colonoscopy results, recruited subjects were classified into three categories: 1) healthy subjects: individuals with colonoscopy negative for tumor, adenomas and other diseases; 2) adenoma patients: individuals with colorectal adenoma/s; and 3)

CRC patients: individuals with newly diagnosed CRC. A total of 80 subjects were recruited divided into 29 CRC patients, 27 adenomas and 24 controls. Stool was collected in Stool Nucleic Acid Collection and Transport Tubes with RNA stabilising solution (Norgen Biotek Corp) and returned before performing the colonoscopy. Aliquots of the stool samples were stored at -80°C until use. DNA was extracted from aliquot of fecal samples using the Qiamp DNA stool kit (Qiagen).

For Cohort2, samples were collected from patients recruited at the European Oncology Institute in Milan, Italy with the help of our collaborators Drs. Maria Rescigno, Chiara Pozzi, Sara Gandini and Davide Serrano. A total of 60 subjects were recruited, divided into 32 CRC patients and 28 controls. Controls, matched for age (± 5 years) and season of blood withdrawn (± 2 years), were recruited among subjects who underwent recent colonoscopy and had negative or no other relevant gastrointestinal disorders. Subjects reporting the use of antibiotics were excluded. Fecal samples were collected from healthy subjects and patients (before surgery, or any other cancer treatment) and directly frozen at -80°C in resuspension buffer (TES buffer: 50 mM Tris-HCL, 10 mM NaCl, 10 mM EDTA, pH 7.5) and kept in liquid nitrogen until DNA extraction. DNA was extracted from fecal samples with G'NOME DNA isolation kit (MP) following a published protocol (Furet et al. 2009).

With the help of Federica Armanini, sequencing libraries were prepared using the NexteraXT DNA Library Preparation Kit (Illumina), following the manufacturer's guidelines. Sequencing was performed on the HiSeq2500 (Illumina) at the internal sequencing facility of the Centre for Integrative Biology, Trento, Italy.

With the help of Paolo Manghi and Edoardo Passoli, we downloaded 5 public fecal shotgun CRC datasets covering samples from 6 different countries, totaling 313 CRC patients, 143 adenomas and 308 controls (**Table 2**). We manually curated metadata tables for the public cohorts according to the curatedMetagenomicDataset (Pasolli et al. 2017) R-package grammatical rules (cMD). The metadata table includes ten fields (sampleID, subjectID, body_site, country, sequencin_platform, PMID, number_reads, number_bases, minimum_read_length, median_read_length) that are mandatory for all datasets in addition to other fields that are dataset-specific.

3.1.2 Sequence pre-processing, taxonomic and functional profiling

With the help of Francesco Asnicar, fecal metagenomic shotgun sequences obtained from the Italian cohorts were subjected to a pre-processing pipeline whereby sequences were quality filtered using trim_galore (parameters: --nextera --stringency 5 --length 75 --quality 20 --max_n 2 --trim-n) discarding all reads with quality less than 20 and shorter than 75 nucleotides and then aligned to the human genome (hg19) and the PhiX genome for human and contaminant DNA removal

using bowtie2 (Langmead and Salzberg 2012). Samples with a high percentage of reads mapping to the human genome and/or smaller than 1Gb were excluded.

Table 2. Sizes and characteristics of the large-scale CRC metagenomic datasets used in this study.

Dataset	Groups	Age	BMI	Gender F/M	Country	# of reads (x 10^9)
ZellerG_2014 (Zeller et al. 2014)	Control (61) Adenoma (42) CRC (53)	60.6 +/- 11.4 63 +/- 9.1 66.8 +/- 10.9	24.7 +/- 3.2 25.9 +/- 4.1 25.5 +/- 5.2	54.1/45.9 28.5/71.5 45.2/54.8	France	9.36
YuJ_2015 (Yu et al. 2015)	Control (54) CRC (74)	61.8 +/- 5.7 66 +/- 10.6	23.5 +/- 3 24 +/- 3.2	38.9/61.1 35.1/64.9	China	7.2
FengQ_2015 (Feng et al. 2015)	Control (61)* Adenoma (47) CRC (46)	67 +/- 6.5 66.5 +/- 7.9 67 +/- 10.9	27.6 +/- 3.8 28 +/- 4.7 26.5 +/- 3.5	41/59 51.1/48.9 39.1/60.9	Austria	8.3
VogtmannE_2016 (Vogtmann et al. 2016)	Control (52) CRC (52)	61.2 +/- 11 61.8 +/- 13.6	25.3 +/- 4.2 24.9 +/- 4.2	28.8/71.2 28.8/71.2	USA	6.9
HanniganGD_2017 (Hannigan et al. 2017)	Control (28) Adenoma (27) CRC (27)	NA	NA	NA	USA (54) Canada (28)	0.53
CM_Cohort1 (This study)	Control (24) Adenoma (27) CRC (29)	67.9 +/- 7.1 62.8 +/- 8.6 71.4 +/- 8.2	25.3 +/- 3.5 25.3 +/- 4.1 25.7 +/- 4.1	45.8/54.1 40.7/59.3 20.7/79.3	Italy	8.18
CM_Cohort2 (This study)	Control (28) CRC (32)	57.8 +/- 8.3 58.4 +/- 8.4	24.6 +/- 3.8 26.8 +/- 4.3	42.9/57.1 28.1/71.9	Italy	5.15
Total	**Control (308)** **Adenoma (143)** **CRC (313)**	--	--	--	--	**45.6**

*Numbers differed from the original sample numbers reported in the article due to metadata and/or sequence processing issues. NA = Not available.

We used MetaPhlAn2 (Truong et al. 2015) for quantitative profiling the taxonomic composition of the microbial communities of all metagenomic samples, whereas HUMANn2 (Abubucker et al. 2012) was used to profile pathway and gene family abundances and StrainPhlAn (Truong et al. 2017) and PanPhlAn (Scholz et al. 2016) for strain-level profiling. Oral species were defined by analyzing 463 oral samples from the Human Microbiome Project dataset (Human Microbiome Project Consortium 2012) and 140 saliva samples from (Brito et al. 2016). Species with >0.1% abundance and >5% prevalence were deemed to be of oral origin.

3.1.3 Machine learning analysis

With the help of Paolo Manghi, our machine learning analyses exploited 4 types of microbiome quantitative profiles: taxonomic species-level relative

abundances and marker presence/absence patterns inferred by MetaPhlAn2 (Truong et al. 2015), gene-family and microbial pathway relative abundances estimated by HUMAnN2 (Abubucker et al. 2012).

All machine learning experiments were run using Random Forest (Breiman 2001), as this approach has been shown to outperform other popular learning tools for microbiome data (Pasolli et al. 2016). Analyses were conducted using MetAML (Pasolli et al. 2016) with the Random Forest implementation taken from Scikit-Learn (Pedregosa et al. 2011). We used an ensemble of 1000 estimator trees and Shannon entropy to evaluate the quality of a split at each node of a tree; we set the minimum number of samples per leaf at 5 and 30% of the total number of features, and chose no maximum depth of the trees, as indicated elsewhere (Hastie et al. 2009). For the marker presence/absence profiles we used the number of features equal to the square root of the total number of features, and this percentage was further decreased to 1% when using gene-family profiles as they have a substantially higher number (>2 million gene families).

We measured the inside-dataset prediction capability through 10-fold cross-validations. Each fold contained a balanced proportion of positive and negative classes. The procedure of forming the folds and assessing the models was repeated 20 times and the final result is an average over 200 tests. In the LODO (Leave-One-Dataset-Out) approach, the moiety of data excluded from the training for testing corresponds to a whole cohort of patients. This technique was applied to all the cohorts, with each cohort hosting a balanced number of positive and negative cases. In the cross-stage prediction, datasets were considered two by two: one was used for training the model, the other to test it.

3.1.4 Statistical analysis

We performed univariate analyses on a per dataset basis using LEfSe (Segata et al. 2011) to identify features that were statistically different among groups and estimate their effect size. Multivariate analysis was conducted using linear models fitted to the data using the limma R package (Ritchie et al. 2015) and possible confounders such as age, gender and BMI were included in the models. For the meta-analysis on taxonomic and functional profiles, we converted relative abundances to arcsine-square root transformed proportions and used the *escalc* function from the R metafor package that employed Hedge's standardized mean difference statistic to calculate random effects model estimates. We quantified study heterogeneity using the I^2 estimate (percentage of variation reflecting true heterogeneity) as well as Cochran's Q test to assess statistically significant heterogeneity. P-values obtained from the random effects models were corrected for multiple hypothesis testing correction using the Benjamini-Hochberg procedure and corrected *p-values* <0.05 were considered statistically significant. Cluster analysis

was conducted by calculating distance matrices from phylogenetic trees using the APE R package, clustering using partitioning around medoids (PAM) and computing clusters' prediction strength using the cluster R package. Associations between distances and categorical metadata was performed using the *anosim* function from the vegan R package.

3.1.5 Identification and quantification of TMA producing enzymes

We downloaded amino acid sequences that matched the keywords "*cutC*" and "*cutD*" from UniProt90 (Apweiler et al. 2004), mapped their IDs to EMBL CDS using UniParc and used the resulting DNA sequences to search, using BLASTn (Altschul et al. 1990), all 48,902 prokka (Seemann 2014) annotated genomes available in our repository (Segata et al. 2013). Matching queries were filtered to include only alignments with >80% identity and length >1000nt for *cutC* and >800nt for *cutD*, and a e-value <1e-15. We used ShortBRED (Kaminski et al. 2015) to identify short seed sequences that were representative of the filtered queries using UniProt's UniRef100 database and quantified them in the metagenomes, normalizing by the number of reads per kilobase million (RPKM). The pipeline was also applied to identify and quantify the L-carnitine/gamma-butyrobetaine antiporter (*caiT*) and the dioxygenase *yeaW*, responsible for producing TMA preferentially via carnitine degradation (Koeth et al. 2014). In order to investigate differences in *cutC* sequence types, we clustered *cutC* sequences at 97% sequence identity using UCLUST (Edgar 2010) and aligned raw reads to the clustered *cutC* database using bowtie2 (Langmead and Salzberg 2012). From the bam files we calculated the breadth and depth of each sequence and generated their corresponding consensus sequence using Samtools (Li et al. 2009) and VCF utils (Danecek et al. 2011). We chose the representative *cutC* sequence for each sample using two criteria: i) the highest breadth; ii) if there were multiple *cutC* sequences with the same breadth, the one with highest depth. We filtered representative *cutC* sequences from each sample to include only those with a breadth >80%, aligned them using MAFFT (Katoh and Standley 2013), built a phylogenetic tree using fastTree (Price, Dehal, and Arkin 2010) which was refined using RAxML (Stamatakis 2014) and visualized using GraPhlAn (Asnicar et al. 2015).

3.2 Results

3.2.1 Italian cohorts

We performed shotgun metagenomic sequencing of the stool microbiome of 140 CRC patients and controls recruited in two distinct italian cohorts and obtained a total of 4.1 x 10^9 paired-end sequences, with an average depth of 23 million ± 12 million human DNA-free metagenomic reads.

Quantitative microbiome taxonomic profiles revealed microbial species significantly associated with CRC for both of the newly sequenced cohorts (**Figure 7A-B**). Several of these associations were however dataset-specific. Bacterial species enriched in CRC in at least one of the two datasets included *Escherichia coli, Fusobacterium nucleatum, Clostridium symbiosum*, *Parvimonas micra*, *Peptostreptococcus stomatis*, *Eubacterium ventriosum*, *Gemella morbillorum*, *Porphyromonas asaccharolytica* and *Solobacterium moorei*, whose associations with CRC have been already reported (Zeller et al. 2014; Feng et al. 2015; Yu et al. 2017). On the other hand, enrichment of the methanogenic *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* in CRC has not been previously reported in the literature.

Although the adenoma-associated microbiome has been reported to resemble the healthy gut microbiome (Zeller et al. 2014), we found three bacterial species significantly enriched in adenomas. The first two are *Rothia dentocariosa* and *Staphylococcus aureus* and could be consistent with an increased oral-to-gut microbial influx in disease (Flemer et al. 2017; Kostic et al. 2012; Drewes et al. 2017). The third adenoma-associated organism is *Collinsella aerofaciens*, which has been linked with increased CRC risk (Moore and Moore 1995), and also with proinflammatory cytokines such as tumor necrosis factor alpha (*TNF-α*) and the Monocyte Chemoattractant Protein 1 (*MCP-1*) (Zhang et al. 2001; Tanaka et al. 2006). Both the number of microbial species and gene richness were higher in CRC patients compared to controls for Cohort2 (85 vs 73 average species $p = 0.0002$, and 286,141 vs 265,278 gene families $p = 0.01$ - **Figure 7C**), but these differences were not detected in Cohort1.

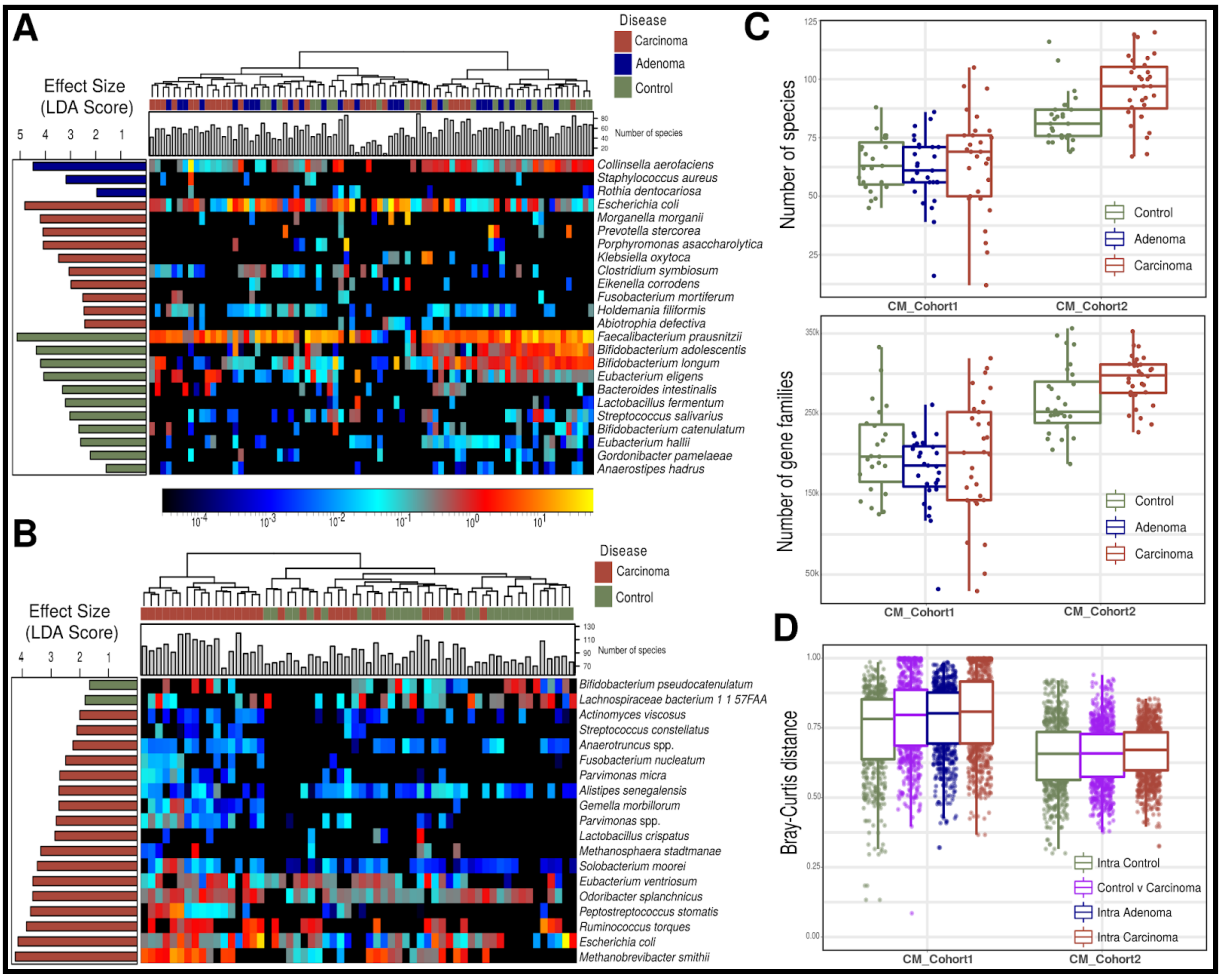


Figure 7 - Two novel metagenomic cohorts identify clear but only partially overlapping microbiome signatures associated with CRC. (a) Relative abundances (log scale) and effect sizes (estimated using the LDA score in LEfSe) for the significantly different microbial species in CRC-associated samples compared to control samples for Cohort1 and **(b)** for Cohort2. **(c)** Alpha diversities measured as the total number of species and the total number of UniProt90 gene families in each sample for the two cohorts. **(d)** Beta-diversities estimated with the Bray-Curtis dissimilarity metric for intra- and inter-condition comparisons in the two cohorts.

3.2.2 Taxonomic and functional profiling of all available datasets

The five publicly available datasets were also sequenced at high depth (from average 43 million ± 18 million to 66 million ± 15 million, **Appendix 5**) except for the Hannigan et al. study (6.5 million ± 3.8 million).

When we evaluated microbial richness and diversity in CRC samples and controls, we found species richness was higher in CRC samples compared to controls, and the increase was significant in four of the six deeply sequenced datasets (p < 0.05) when accounting for differences in read depth and is not confounded by age, BMI or gender. Meta-analysis of standardized mean differences using the number of microbial species by random effects model (DerSimonian and Laird 1986) confirmed the presence of a higher number of bacterial species in CRC compared to controls (μ=0.5, 95% CI [0.16, 0.85], *p-value* = 0.004), although with significant heterogeneity across datasets (I^2 = 74.8%, *p-value* = 0.0007, Q-test).

Shannon diversity, however, was instead neither increased nor decreased between disease conditions (**Appendix 7**). We also tested whether the CRC-associated microbiome possesses more oral cavity-associated species than controls, as previously hypothesized (Flemer et al. 2017; Drewes et al. 2017). Using 153 species identified from existing datasets (Brito et al. 2016; Human Microbiome Project Consortium 2012) as being typical colonizers of the oral cavity, we found increased oral species richness in CRC samples for all six deeply sequenced datasets compared to controls (one significant with *p-value* 0.01; meta-analysis μ=0.16, 95% CI [-0.03, 0.35], *p-value* = 0.017) (**Appendix 8**). Similarly, the total abundance of oral species in the stool microbiome of CRC patients was also significantly higher compared to controls (meta-analysis μ=0.23, 95% CI [0.07, 0.39], *p-value* = 0.003).

Several CRC biomarker species were identified by univariate statistics using LEfSe (Segata et al. 2011) independently in the majority of the datasets: *Fusobacterium nucleatum*, *Solobacterium moorei*, *Porphyromonas asaccharolytica*, *Parvimonas micra*, *Peptostreptococcus stomatis,* and *Parvimonas ssp.* Other species were identified in fewer datasets or were dataset specific (**Figure 8A**). Some of the cross-cohort CRC biomarker species have already been suggested (Flemer et al. 2017; Kostic et al. 2012; Drewes et al. 2017) and many of them are commonly found in the oral cavity (8 out of the 39 total biomarkers found in at least 2 datasets) consistently with the increased oral taxa presence in CRC samples mentioned above.

We then pooled evidence of differential abundance across datasets by random effects meta-analysis. Among the 21 differentially abundant species at FDR < 0.005, those with the highest effect size were again *F. nucleatum*, *S. moorei*, *P. asaccharolytica*, *P. micra* and *P. stomatis*. The meta-analysis additionally identified *Clostridium symbiosum*, which has been already tested as a marker for early CRC detection (Xie et al. 2017) (**Figure 8B**). Other differentially abundant species at FDR < 0.05 have not been previously reported in CRC case/control studies, including *Streptococcus tigurinus* and *Streptococcus dysgalactiae*, and 3 different *Campylobacter* species (**Appendix 9)**. *Campylobacter* has been shown to co-occur with *Fusobacterium* species, pointing at a potential synergistic effect of such microbes (Warren et al. 2013). We also found *Gemella morbillorum* and *Streptococcus gallolyticus* to be relevant biomarkers, as previously suggested in smaller cohorts (Feng et al. 2015; Andres-Franch et al. 2017; Boleij et al. 2011). In contrast, only 12 species were associated with the control population in the meta-analysis and only four were significantly enriched for the same populations in at least 3 datasets. Control-associated species with the highest effect sizes were *Gordonibacter pamelae* and *Bifidobacterium catenulatum* (**Figure 8B**), which are generally considered beneficial microbes and have been used as probiotic supplements (Picard et al. 2005; Fijan 2014). The substantially higher number of species enriched in CRC than in controls (49 vs 12), even when focusing only on

species with putative oral origin (16 vs 2, **Appendix 10**), points to the existence of a reproducible taxonomic signature of the CRC-associated microbiome and a generally higher relative abundance of non-cancer associated species that are much more diverse and thus harder to be identified through differential abundance analysis.

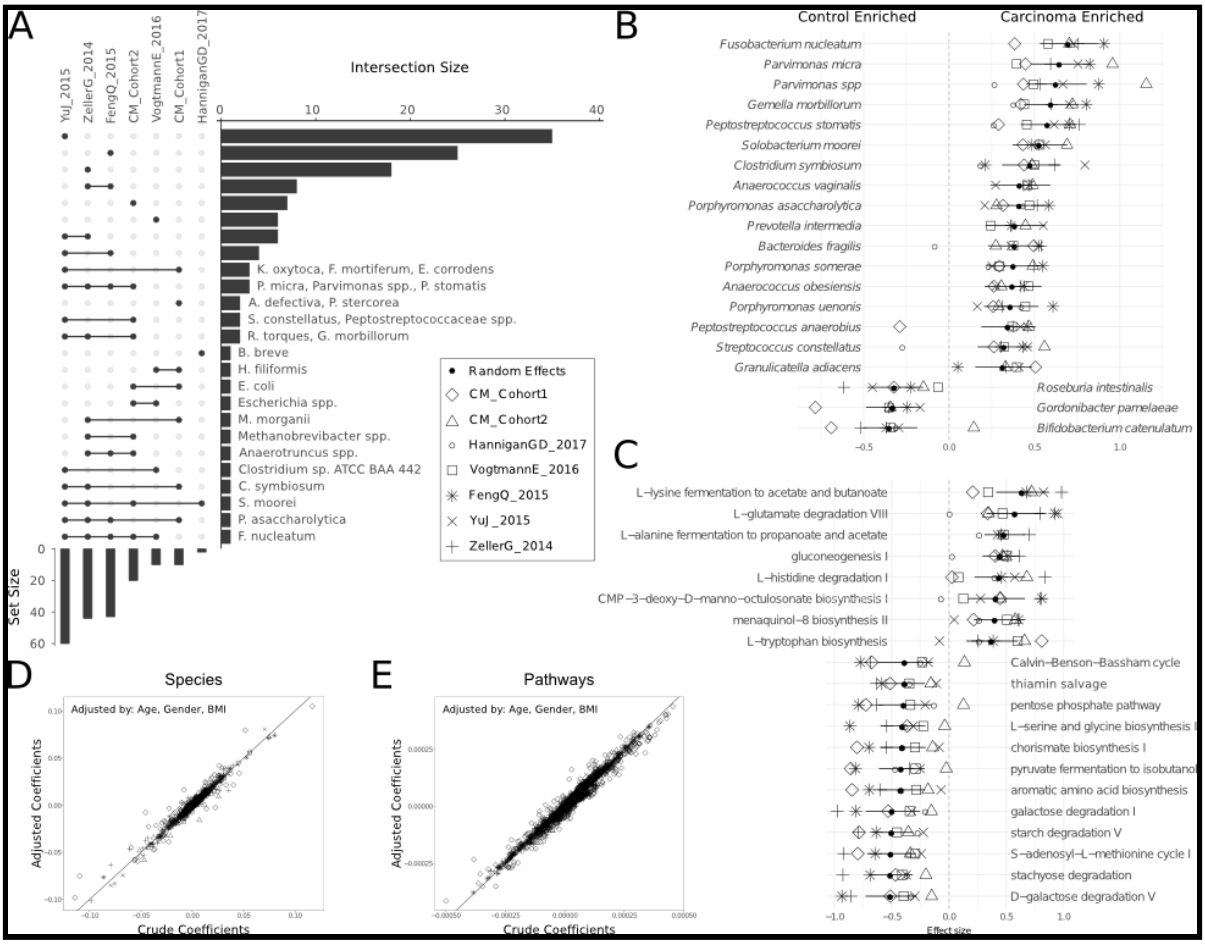


Figure 8 - Reproducible taxonomic and functional microbial biomarkers for CRC across datasets. (a) Upset plot showing the number of taxonomic biomarkers identified using LefSE on MetaPhlAn2 species profiles shared by combinations of datasets. **(b)** Pooled effect sizes for the 20 significant features with the largest effect size calculated using a meta-analysis of standardized mean differences and a random effects model on MetaPhlAn2 species abundances and on **(c)** HUMANn2 pathway abundances**.** Bold lines represent the 95% confidence interval for the random effects model coefficient estimate (marked with a black circle). **(d)** Scatter plot of crude and age, gender and BMI adjusted coefficients obtained from limma for MetaPhlAn2 species abundances. **(e)** Scatter plot of crude and age, gender and BMI adjusted coefficients obtained from limma for HUMANn2 pathway abundances.

Functional potential of the microbiome was also significantly associated with CRC samples. We found overall increased richness of UniProt gene families in CRC samples in two of the 7 datasets, with percentages of unmapped reads ranging between 20 and 40% (**Appendix 11**). We found 35,271 single gene families associated with CRC and 23,562 associated with controls (FDR < 0.05). Of the metagenomically reconstructed whole microbial pathways, 175 were associated with disease status, divided in 164 CRC-associated and only 11 control-associated pathways. Among pathways with the largest effect size (**Figure 8C**), we found

starch, stachyose and galactose degradation to be associated with control status. We also found that the CRC-associated microbiota showed an association with gluconeogenesis and with the capacity for uptake and metabolism of amino acids via the putrefaction (L−histidine degradation to L-glutamate) and fermentation pathways.

Using strain-level resolution of species based on single nucleotide variants, we found geographically distinct *Eubacterium rectale* and *Dorea formicigenerans* strains between westernized and non-westernized samples (**Figure 9A**), in concordance with previously published results (Truong et al. 2017), however, we found no apparent disease-specific strains (**Appendix 12**). Using strain-level resolution by identifying gene composition, we reconstructed 2 *Bacteroides fragilis* sub-species present in 109 samples (**Figure 9B**) and found a subtree to be marginally enriched with CRC samples (p = 0.08, Fisher Test).

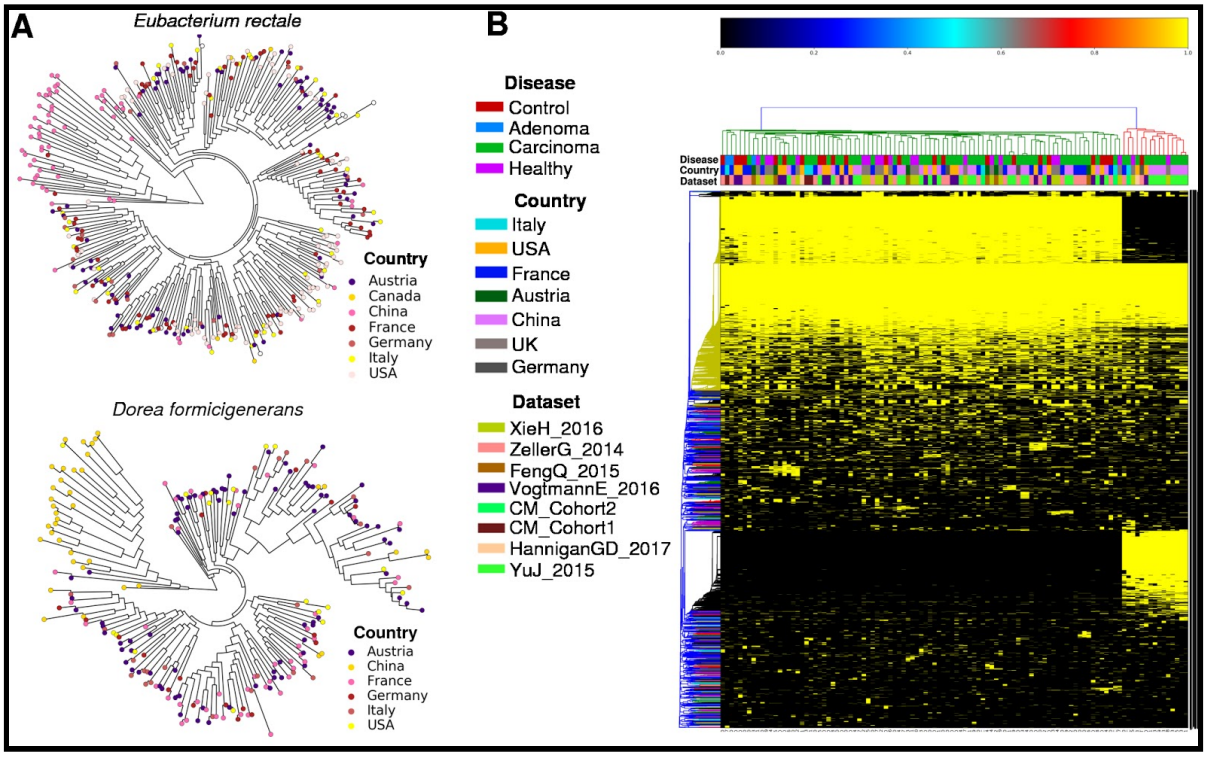


Figure 9 - Complementary strain-level analyses reveal associations with geography and disease. **(a)** Strain population structures for *Eubacterium rectale* and *Dorea formicigenerans* associated with geography, reported as phylogenies built on the concatenated alignments of each species-specific reconstructed marker set. **(b)** Gene family presence/absence profiles of 109 metagenomically detected *Bacteroides fragilis* strains using 58 reference genomes. Each column represents a strain from a sample and each row a gene family.

3.2.3 Machine learning analysis

When assessing the predictability of the CRC condition using quantitative species-level taxonomic profiling in each cohort via unbiased cross validation, we observed performances ranging from 0.59 AUC score for the Hannigan et al. dataset (Hannigan et al. 2017) to 0.92 AUC score for the Feng et al. dataset (Feng et al. 2015) (average 0.80 AUC, **Figure 10A**). The dataset with the worst AUC score was

sequenced at the lowest depth (9.5 times fewer reads per sample than the mean), which may explain its poor prediction accuracy. When we considered the functional potential of the gut microbiome by means of UniProt90 gene family abundances, we found that predictions improved, with AUCs ranging from 0.49 (Hannigan et al. dataset) to 0.83 (Zeller et al. dataset - with a mean of 0.75, **Figure 10B**).

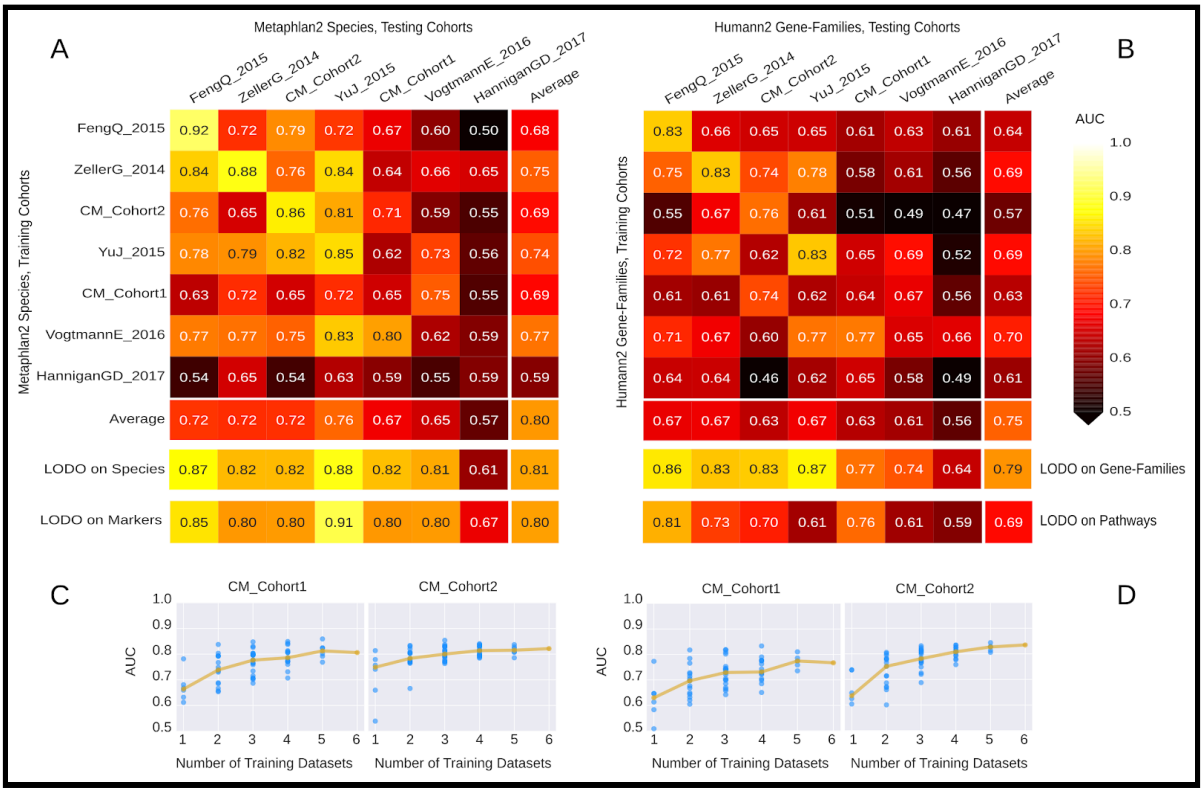


Figure 10 - Assessment of prediction performances of the gut microbiome for CRC detection within and across cohorts. **(a)** Cross prediction matrix reporting prediction performances as AUC values obtained using a random forest (RF) model on species-level relative abundances. Values on the diagonal refer to 20 times repeated 10-fold stratified cross validations. Off-diagonal values refer to the AUC values obtained by training the classifier on the dataset of the corresponding row and applying it on the dataset of the corresponding column. The Leave-One-Dataset-Out (LODO) row refers to the performances obtained by training the model on the species-level abundances and MetaPhlAn2 markers using all but the dataset of the corresponding column and applying it on the dataset of the corresponding column. (**b)** Cross prediction matrix of AUC values obtained using HUMANn2 UniProt90 gene-family abundances and HUMANn2 pathway relative abundances. **(c)** Prediction performances for the two Italian cohorts at increasing numbers of external datasets considered for training the model. The dark yellow line interpolates the median AUC at each number of training datasets considered. **(d)** Prediction performances at increasing number of datasets in the training, using HUMANn2 UniProt90 gene-family abundances.

We then tested whether and how much the microbial signatures of CRC remained predictive across distinct datasets and cohorts. To this end, we trained the random forest classifier on each single “training” dataset and applied the model on each distinct “testing” dataset. For most of the datasets this led to decreased AUC values when compared to single cross validation AUCs, maintaining a significant AUC spread across cohorts (minimum 0.5 and maximum 0.84 cross dataset AUC). These results were further confirmed when we used either pathway or gene family-abundances for CRC prediction (**Figure 10B**). For datasets whose CV AUC values were greater than 0.8 (4 datasets), this approach decreased prediction power,

whereas datasets with poorer AUC values (3 datasets) achieved slight improvements in prediction. A feature selection approach was also applied by exploiting the internal feature ranking of the Random Forest classifier. These results are presented in Figure 3 of Thomas, Manghi et al. 2019.

To try to overcome the limitations of single cross-dataset predictions, we performed a Leave-One-Dataset-Out (LODO) analysis (Riester et al. 2014) in which the microbiome-CRC association of the samples in each dataset was predicted using a classifier trained using the combination of all samples in the remaining datasets. For taxonomic profiles, we found this approach to work best in terms of prediction accuracies, reaching AUC values >0.80 for all the six deeply sequenced datasets (**Figure 10A**). For functional profiles, gene families performed similarly well, with AUC values >0.75, whereas pathway abundances were not comparably informative in predicting CRC (**Figure 10B** - Bottom).

When we assessed the prediction performances using increasingly larger subsets in the training cohort, AUC values sharply increased when moving from one to two training datasets (7% median AUC improvement) and from two to three datasets (2% median AUC improvement, **Appendix 13**). Prediction performances tended to plateau for some of the testing datasets when using the largest possible training set (**Figure 10C-D**), but improvements (>1% AUC) were still detected for three datasets even when transitioning from using 5 to 6 training datasets.

3.2.4 Quantification of choline TMA-lyase enzymes

Microbiome-derived metabolites have been implicated in carcinogenesis (Di Martino et al. 2013; Ou et al. 2012). We chose to focus on trimethylamine (TMA), an amine produced by bacteria from choline and carnitine, because it has been recently shown to play a role in complex diseases such as atherosclerosis (Jie et al. 2017), which shares molecular pathways of disease development and progression with cancer (Ross et al. 2001), and primary sclerosing cholangitis (Kummen et al. 2017). Since dietary components have been shown to be linked with CRC risk (Johnson et al. 2013; Wei et al. 2004; Huxley et al. 2009), we hypothesized that the TMA-producing potential of the human gut microbiome could also be associated to CRC (Oellgaard et al. 2017). To test this hypothesis, we built a database for genes belonging to the main TMA-synthesis pathways and then used them to reconstruct and quantify the presence of such genes in the 764 CRC-associated metagenomes. The main genes associated with TMA-synthesis are those encoding for the choline TMA-lyase (*cutC*), the L-carnitine dioxygenase (*yeaW*) and the L-carnitine/gamma-butyrobetaine antiporter (*caiT*) and we identified them in 923, 5185 and 5709 available bacterial genomes, respectively. Putative *cutC* sequences belonged mainly to *Proteobacteria* (mostly *Gamma*- and few *Deltaproteobacteria*) and *Firmicutes* (mainly from *Clostridia* and few *Bacilli*), with few *Actinobacteria* as

reported previously (Rath et al. 2017). Screening the 7 CRC-associated metagenomic datasets, we found that only one dataset had a significant increase of *caiT* in CRC samples compared to controls, whereas no significant differences were detected for *yeaW*. However, we found increased abundance of *cutC* in CRC samples compared to controls in all seven datasets, with statistical significance at alpha 0.05 for 4 datasets (**Figure 11A**) and a significant *p-value* (0.002) when meta-analysing all samples together (μ=0.23 with 95% CI: 0.08 to 0.39, I^2 = 0%, Q-test = 0.7).

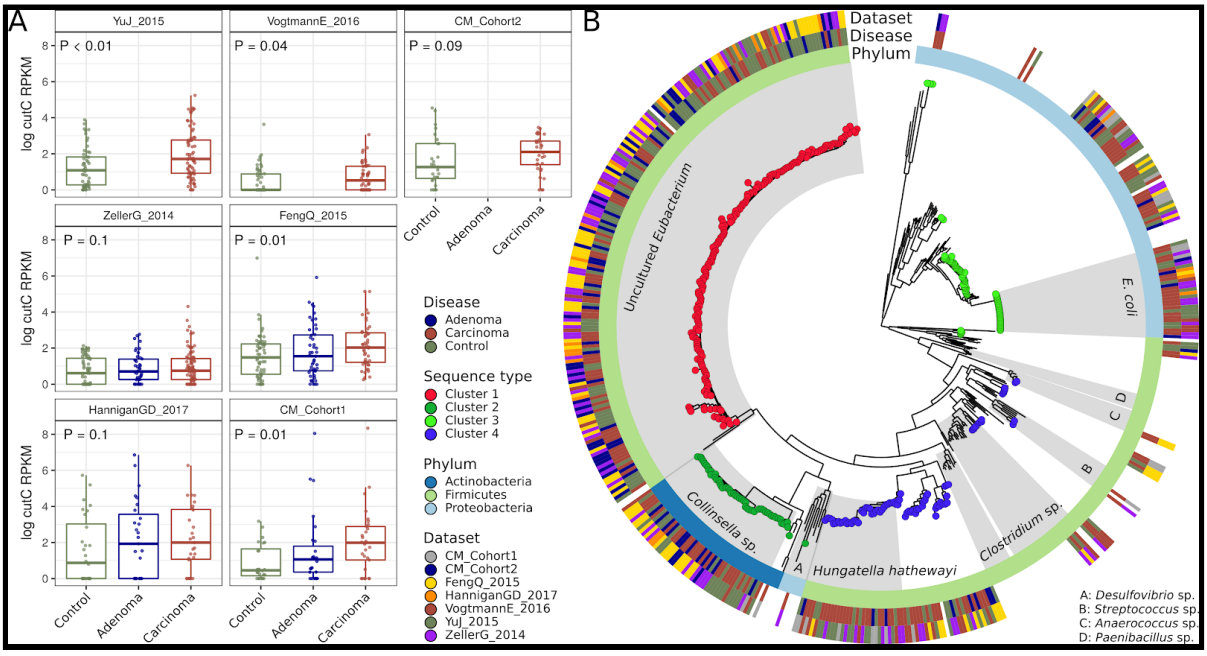


Figure 11 - Choline TMA-lyase *cutC* and its genetic variants are a strong biomarker for CRC-associated stool samples. (a) Boxplots showing the log of reads per kilobase million (RKPM) abundances obtained using ShortBRED for the choline TMA-lyase enzyme *cutC*. P-values were calculated by wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **(b)** Phylogenetic tree of sample-specific cutC sequence variants identifies four main sequence subtypes. Tips with no circles represent *cutC* sequence variants from genomes absent from the datasets. Taxonomy was assigned using BLASTn and a *cutC* sequence database (criteria of 80% coverage, >97% identity and minimum 2000 nt alignment length).

We further explored the role of *cutC* in the gut microbiome by reconstructing sample-specific sequence variants using a reference-aided targeted assembly approach. We found a large sequence divergence for the gene encoding this enzyme and identified four main sequence subtypes that correlated with taxonomy. The most prevalent (46.5%) *cutC* sequence type belonged to an unknown uncultured *Eubacterium* species with only 95% sequence identity to the closest known and taxonomically characterized variant. This *cutC* variant was associated with non-CRC samples (OR 0.38, 95% CI [0.25, 0.57], p = 0.0001, Fisher Test), whereas *cutC* sequence types mostly belonging to *Hungatella hathewayi and Clostridium asparagiforme* (*Firmicutes*) were significantly CRC-associated (OR 2.14, 95% CI [1.29, 3.56], p = 0.004, Fisher Test), as were sequence types belonging to *Klebsiella oxytoca* and *Escherichia coli* (OR 1.85, 95% CI [1.13, 3], p = 0.02, Fisher Test - **Figure 11B)**.

3.3 Discussion

This study was performed across multiple datasets, populations, and conditions, through a combined analysis of fecal CRC metagenomes from two previously unpublished cohorts and five publicly available datasets. Robustly linking gut microbiome members and functions with CRC samples is currently a crucial task because although direct specific host-microbe interactions have been shown to be a causal link for malignancies (Dejea et al. 2018; Cougnoux et al. 2014; Wu et al. 2009; Chung et al. 2018), indirect mechanisms may have a higher impact on the development of carcinomas. In our analysis of seven independent cohorts, we indeed found a panel of reproducible microbiome members as well as strain-level determinants beyond the validated mechanisms in specific variants of *Escherichia coli* (Cougnoux et al. 2014; Dejea et al. 2018; Cuevas-Ramos et al. 2010) or *Bacteroides fragilis* (Dejea et al. 2018; Ulger Toprak et al. 2006). In addition to previously identified bacteria such as *F. nucleatum*, *Parvimonas*, and *Peptostreptococcus stomatis,* whose association we consistently confirm, newly identified correlates include *Streptococcus tigurinus*, *Streptococcus dysgalactiae*, and *Campylobacter* species. We find that the gut microbiome in CRC has greater richness than controls, partially due to the presence of oral cavity-associated species rarely found in healthy guts, challenging the hypothesis that decreased alpha-diversity is always associated with intestinal dysbiosis.

The identification of reproducible microbial biomarkers for CRC may enable the design of non-invasive diagnostic tools. We developed machine learning models able to distinguish between carcinoma patients and controls with an average performance above 0.81 AUC when validated on fully independent datasets. Larger datasets (n > 1,000) will very likely further increase this performance and the combination of a microbiome model with a fecal occult blood test and patient risk factors could substantially improve this diagnostic accuracy. The accuracy of CRC prediction increases with the number of microbes or microbial genes used, with single biomarkers being much inferior to multi-featured diagnostic models. However, nearly maximal accuracy was achieved with as few as 15 to 25 microbes or few hundred genes, providing the potential for an inexpensive clinical microbiological test.

Methodological aspects are critical to the evaluation of the gut microbiome as a potential diagnostic tool. Datasets can be sensitive to technical artefacts arising from different experimental procedures, such as the choice of DNA extraction, which causes significant effects on the observed microbial profiles (Costea et al. 2017), and to heterogeneity of factors implicated in microbial shifts in healthy populations, including biogeography, diet, and host genetics (David et al. 2014; Bonder et al. 2016; Truong et al. 2017). Rigorously performed cross-study validation, such as

those performed here in a LODO approach (Pasolli et al. 2016), provides a realistic expectation of prediction accuracy for cohorts with a similar proportion of disease and controls. First, in contrast to the case-control cohorts analyzed here, any potential target population of a diagnostic test will have much lower disease prevalence. Relevant accuracy measures of a test score, such as Positive Predictive Value, can however be calculated for populations of various disease prevalence given sensitivity and specificity estimated from case-control studies (Vollmer 2006). Second, pitfalls such as using all patients to identify discriminative features before performing cross-validation, can be highly optimistically biased and hence should never be reported (Simon et al. 2003). Recent reports of microbiome-based CRC prediction accuracy (Dai et al. 2018) are unfortunately affected by this pitfall. We thus reiterate the importance of rigorously tested models to provide a realistic estimate of the potential of microbiome-based diagnostic tools.

The diversity and subject-specificity of the human gut microbiome is not yet fully uncovered, with many microbial genes having unknown function, and with strain-level diversity that is transparent to current analysis pipelines. Large scale shotgun metagenomics can begin to overcome this, as shown here by the novel identification of a link between CRC and the microbial pathway producing trimethylamine from choline (Kalnins et al. 2015; Craciun and Balskus 2012; Craciun et al. 2014). The key enzyme for this pathway, the *cutC* choline TMA-lyase, is more abundant in the gut microbiomes of adenoma patients and further elevated in carcinoma patients, with specific variants of *cutC* characterizing controls, adenomas, and carcinomas. TMA-producing choline lyases have been found to be associated with atherosclerosis (Jie et al. 2017), and higher plasma trimethylamine oxide (TMAO) and choline levels have been reported to be correlated with CRC risk (Bae et al. 2014; Xu et al. Li 2015; Guertin et al. 2017). We highlighted the importance of strain-level gene resolution in understanding any potential carcinogenic role of *cutC*: CRC-associated variants mostly originated from *Hungatella hathewayi*, *Clostridium asparagiforme*, *Klebsiella oxytoca,* and *Escherichia coli*, whereas no significant enrichment was detected for a *cutC* variant carried by a newly reconstructed *Eubacterium* species. Thus, genetic variants in key microbial genes involved in choline-induced TMA production by the gut microbiome are a plausible and novel potential mechanism for colorectal carcinogenesis. Further work is needed to establish the changes in protein structure and function associated with the genetic variants identified here.

Chapter 4. Neoadjuvant chemotherapy treatment in gastric cancer patients reveals shifts in gastric microbial communities

Gastric carcinomas account for the second highest cancer mortality rate worldwide, with a 5 year survival rate of only 20-30%. The principal line of treatment for these tumors is through neoadjuvant chemotherapy, which aims to eradicate the tumor through the combination of radiotherapy and chemotherapy. However, chemotherapy response is heterogeneous, with no standard methods to assess its efficacy. The gut microbiota has been shown to modulate the host's response to chemotherapeutic drugs and could provide a novel mechanism to explain such heterogeneity. This chapter describes the methods and results used to profile microbial communities of gastric cancer patients who underwent neoadjuvant chemotherapy treatment, aiming to identify microbial markers that could predict treatment efficacy.

4.1 Materials and Methods

4.1.1 Cohort

This work was part of a larger study that investigated epidemiological and genomic aspects of gastric cancers in a sample of the Brazilian population and was funded by FAPESP (2014/26897-0). With the help of Lais Senda and Dr. Thais Bartelli, a total of 36 patients were included after approval by A.C.Camargo Cancer Center's ethics committee (protocols 2134/2015 and 2169/2016).

Patients diagnosed with gastric adenocarcinomas and subjected to neoadjuvant therapy were enrolled in the study, whereas patients who reported the use of antibiotics for at least 4 weeks prior to sample-collection were excluded. Gastric wash samples were collected during endoscopy at AC Camargo Cancer Center's Department of Endoscopy with the help of Dr. Adriane Pelosof and Lais Senda and kept at -20^{o}C until sample processing. Gastric wash pH was measured using pH strips. Gastric juice or gastric wash samples were collected in two time points: i) before the 1st infusion of chemotherapy and ii) on average 16 days after the end of the neoadjuvant treatment. The time interval between both collections points was on average 102 +/- 22 days. Clinical response to neoadjuvant chemotherapy was assessed by a trained medical physician using the percentage of viable tumor cells after treatment and pathological tumor staging after treatment, as a means to access possible 'downstaging'.

4.1.2 DNA extraction

DNA extraction started after incubating the samples for 18 hours in 600µl of a lysis buffer (Qiagen) and 15µl of proteinase K (20µg/µl) at 55ºC. After this period, DNA samples were extracted using a standard phenol chloroform protocol, followed by ethanol precipitation, quantification using a spectrophotometer (Nanodrop – Thermo Scientific).

4.1.3 PCR amplification and sequencing of the 16S rRNA gene

The V3-V4 region of the 16S rRNA gene was amplified by polymerase chain reaction using the primers U341F 5'-CCTACGGGRSGCAGCAG-3' and 806R 5'-GGACTACHVGGGTWTCTAAT-3'. Amplification reactions were carried out in triplicates using Kapa Hotstart High Fidelity Master Mix (Kapa Technologies) and ~20ng of genomic DNA (gDNA) as the PCR template. Thermocycling conditions were: 95°C for 5 min, 25 cycles of 95°C for 45s, 55°C for 30s and 72°C for 45s and a final extension of 72°C for 2 min. Amplicons were verified in agarose gels and between 50-100ng of PCR products were sent for sequencing. Paired-end amplicon sequencing was performed by Neoprospecta Microbiome Technologies (Santa Catarina, Brazil) on an Illumina MiSeq platform using the MiSeq Reagent Kit v3 (600-cycle).

4.1.4 Data processing

16S rRNA primers were trimmed from demultiplexed sequences using the cutadapt (Martin 2011) plugin available in *Qiime2* (v 2018.8.0) (Bolyen et al. 2018) allowing a maximum of 10% mismatches. Resulting sequences were joined using the VSEARCH (Rognes et al. 2016) plugin and filtered by allowing at most 3 nucleotides in direct succession with a phred score less than 10 before being truncated. Sequences that maintained at least 75% of their length following truncation were retained and used as input into the Deblur plugin (Amir et al. 2017). Deblur generates single-nucleotide resolution OTUs (100% sequence identity) after correcting for Illumina sequencing errors and therefore circumvents establishing sequence identity thresholds. This approach results in amplicon sequence variants (ASVs), also known as exact sequence variants (ESVs), oligotypes, zero-radius OTUs (ZOTUs), and sub-OTU (sOTUs). The minimum reads-option was set to 3 and all sequences were trimmed to 400 bp. Additional filtering steps in Deblur included filtering reads which contained PhiX or adapter sequences, only retaining sequences matching to known 16S sequences (greengenes 88% identity OTUs) and filtering PCR-originated chimeras (Amir et al. 2017). The RDP classifier (v 2.12) (Wang et al. 2007) was used to assign taxonomic ranks using a minimum confidence value of

50%. Species level classification was performed using SPINGO (v 1.3) (Allard et al. 2015) trained on release 11.2 of the RDP database and a minimum confidence value of 50%. Species-level assignments were appended to the RDP classification only for species whose genera level assignments were in agreement. Samples were considered to be HP positive from 16S rRNA amplicon sequencing the relative abundance of ASVs classified as HP were >1% of the total.

4.1.5 Alpha and beta diversity analysis

To account for uneven library sizes, we rarefied the ASV table to 1,126 sequences/sample in order to calculate species diversity, using the Shannon-Weaver index (Shannon 1948), the Simpson index (Simpson 1949) and Faith's phylogenetic diversity (Faith and Baker 2007), and richness (by using the observed species) using *qiime2*.

For beta diversity analysis, ASV-representative sequences were aligned using *MAFFT* (Katoh and Standley 2013) against the aligned *greengenes* core set (DeSantis et al. 2006) with *qiime2* default parameters, and the alignments were lanemask-filtered (Lane 1991). A phylogenetic tree was built using FastTree (Price et al. 2009), weighted and unweighted UniFrac (Lozupone and Knight 2005) distances were calculated and a distance matrix was generated.

4.1.6 Statistical analysis

Paired Wilcoxon tests were used to compare mean differences between pre and post chemotherapy patients for genera and ASV abundances. P-values were corrected for multiple testing using the Benjamini and Hochberg procedure (Benjamini and Hochberg 1995). We used ANOSIM and ADONIS (Oksanen et al. 2018) to compare differences in beta-diversity for different clinical variables using 3 distance metrics; weighted UniFrac, unweighted UniFrac and Bray-Curtis for categorical and numerical variables, respectively. Unless otherwise stated, values were reported as mean ± standard deviation and *p-values* <0.05 were considered statistically significant. All calculations were performed within the R statistical computing environment (R Core Team 2018) unless otherwise stated.

4.2 Results

4.2.1 Patient characteristics

We analyzed bacterial communities from gastric wash samples from 72 samples collected from 36 patients before and after neoadjuvant chemotherapy treatment using 16S rRNA high throughput amplicon sequencing (**Table 3**).

Table 3 – Characteristics of gastric adenocarcinoma patients

Demographic	**Patients with gastric adenocarcinomas (n = 36)**
Age, average	58.3 +/- 11.3
Gender, *N* (%) Male Female	 24 (66.7) 12 (33.3)
BMI, average	26.1 +/- 3.9
Lauren's histological classification, *N* (%) Diffuse Intestinal Mixed NA	 18 (50) 11 (30.6) 4 (11.1) 3 (8.3)
H. *pylori*, % Positive Negative NA	 6 (16.7) 17 (47.2) 13 (36.1)
Neoadjuvant chemotherapy treatment, *N* (%) Folfox/Xelox cisplatin/5-fluorouracil + capecitabine/cisplatin epirubicin/oxaliplatin/capecitabine epirubicin/cisplatin/capecitabine docetaxel/cisplatin/5-fluorouracil docetaxel/cisplatin/5-fluorouracil modified	 24 (66.7) 1 (2.7) 4 (11.1) 5 (13.9) 1 (2.8) 1 (2.8)
Pathological tumor staging after treatment, *N* (%) T1-T2 T3-T4 NA	 20 (55.5) 11 (30.6) 5 (13.9)
Vital status, *N* (%) Deceased (cancer) Deceased (other cause) Alive with disease Alive without disease	 5 (13.9) 2 (5.6) 9 (25) 20 (55.5)
Current use of proton pump inhibitors (PPIs - omeprazol, ranitidine, pantoprazol or nexium) Yes No NA	 16 (44.4) 15 (41.7) 5 (13.9)
Complete pathological response after treatment, *N* (%) Yes No	 5 (13.9) 31 (86.1)
Clinical response after treatment, *N* (%) Yes No	 7(19.4) 7 (19.4)

*NA - Not available

4.2.2 Alpha and beta diversity

A total of 5,412,729 sequence reads were generated and after quality filtering, primer trimming, read merging and the Deblur pipeline, 944,010 (17.4%) sequences remained, with an average of 13,111 sequences/sample. When all samples were considered, a total of 3,289 ASVs were obtained.

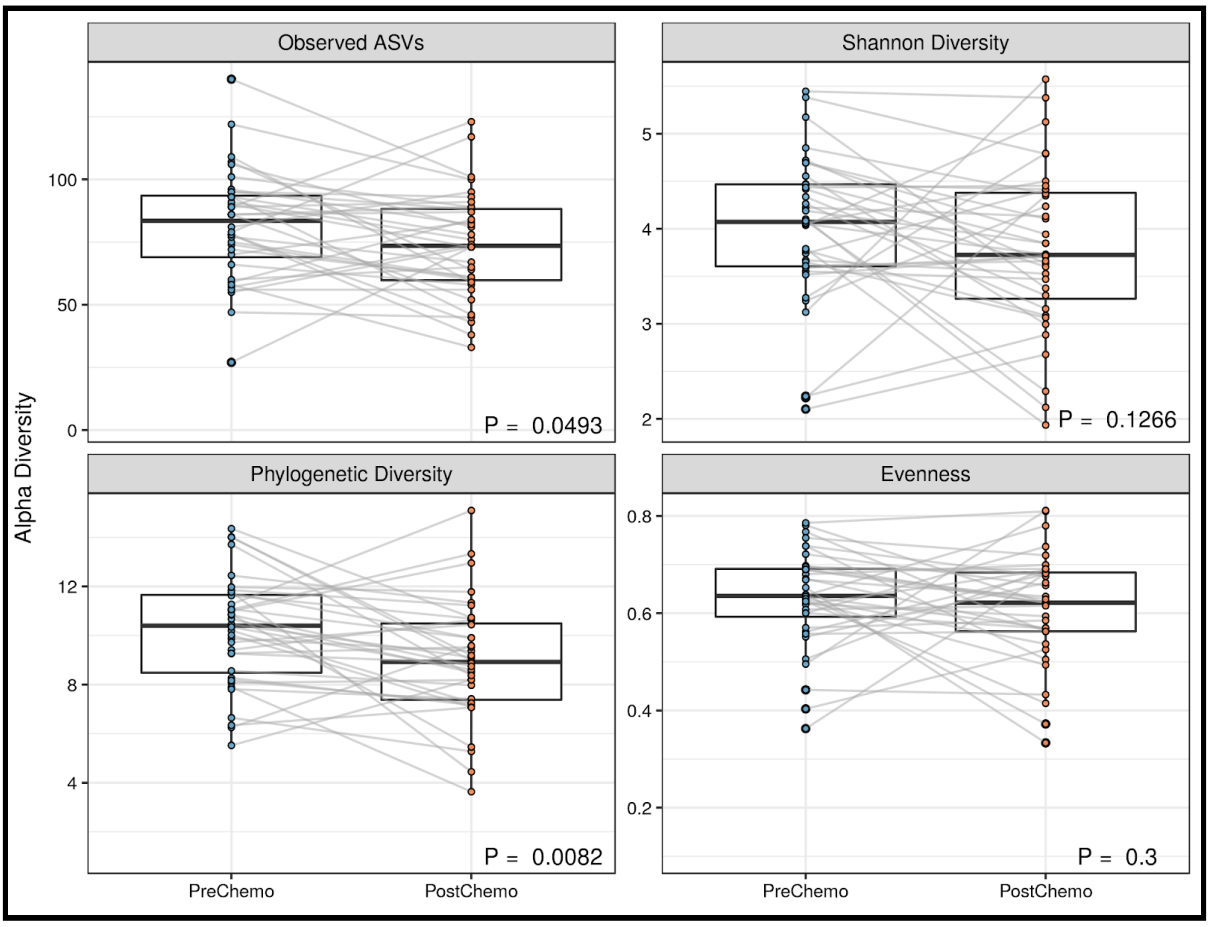


Figure 12 - Alpha diversity before and after chemotherapy treatment. Boxplots showing alpha diversity in gastric adenocarcinoma samples before and after treatment using different metrics (Observed ASVs, Shannon index, Phylogenetic Diversity and Evenness - Pielou's E). Points represent individual samples and lines connecting points represent values for the same patient before and after treatment.

We found a significant decrease in number of observed ASVs and phylogenetic diversity after neoadjuvant chemotherapy ($p < 0.05$, **Figure 12**). However, patients presented mixed directions in terms of alpha diversity, with 21-26 patients presenting decreases and 10-15 patients presenting increases after treatment, depending on the metric.

We used three distance metrics to evaluate differences in microbial communities between different clinical variables (**Table 4**). When considering all three metrics (Bray-Curtis, Unweighted and Weighted UniFrac), we found consistent and statistically significant associations between patients and pH ($p < 0.05$, 999

Permutations), whereas other variables such as pathological response after treatment and sample collection were significant for at most one of the three metrics.

Table 4 – ANOSIM and ADONIS p-values for beta diversity metrics

Variable	Bray-Curtis	Weighted UniFrac	Unweighted UniFrac
Pathological response, pre chemotherapy (Responders = 5) (Non responders = 31)	0.34	0.34	0.37
Pathological response, post chemotherapy (Responders = 5) (Non responders = 31)	0.08	0.15	**0.01**
Clinical response, pre chemotherapy (Responders = 7) (Non responders = 7)	0.73	0.62	0.55
Clinical response, post chemotherapy (Responders = 7) (Non responders = 7)	0.26	0.44	0.58
HP presence, pre chemotherapy (HP positive = 12) (HP negative = 24)	0.47	0.41	0.74
Sample collection, pre or post chemotherapy (N = 72)	0.09	0.06	**0.03**
Gender, pre chemotherapy (Male = 24) (Female = 12)	0.34	0.1	0.58
Age, pre chemotherapy (N = 36)	0.3	0.65	0.44
BMI, pre chemotherapy (N = 36)	0.31	0.75	0.41
Current PPI use, pre chemotherapy Yes (N = 16) No (N = 15)	0.56	0.61	0.44
pH, pre or post chemotherapy (n = 71)	**0.001**	**0.001**	**0.001**
Patient (N = 36)	**0.001**	**0.001**	**0.001**

4.2.3 Changes in microbial abundances in response to treatment

The most abundant genera found in patients before or after chemotherapy treatment included *Streptococcus*, *Prevotella*, *Rothia* and *Veillonella* (which together accounted for an average of 62% relative abundance - **Figure 13**). After multiple hypothesis testing correction, we found no genus to be differentially abundant when comparing patient samples before and after chemotherapy treatment.

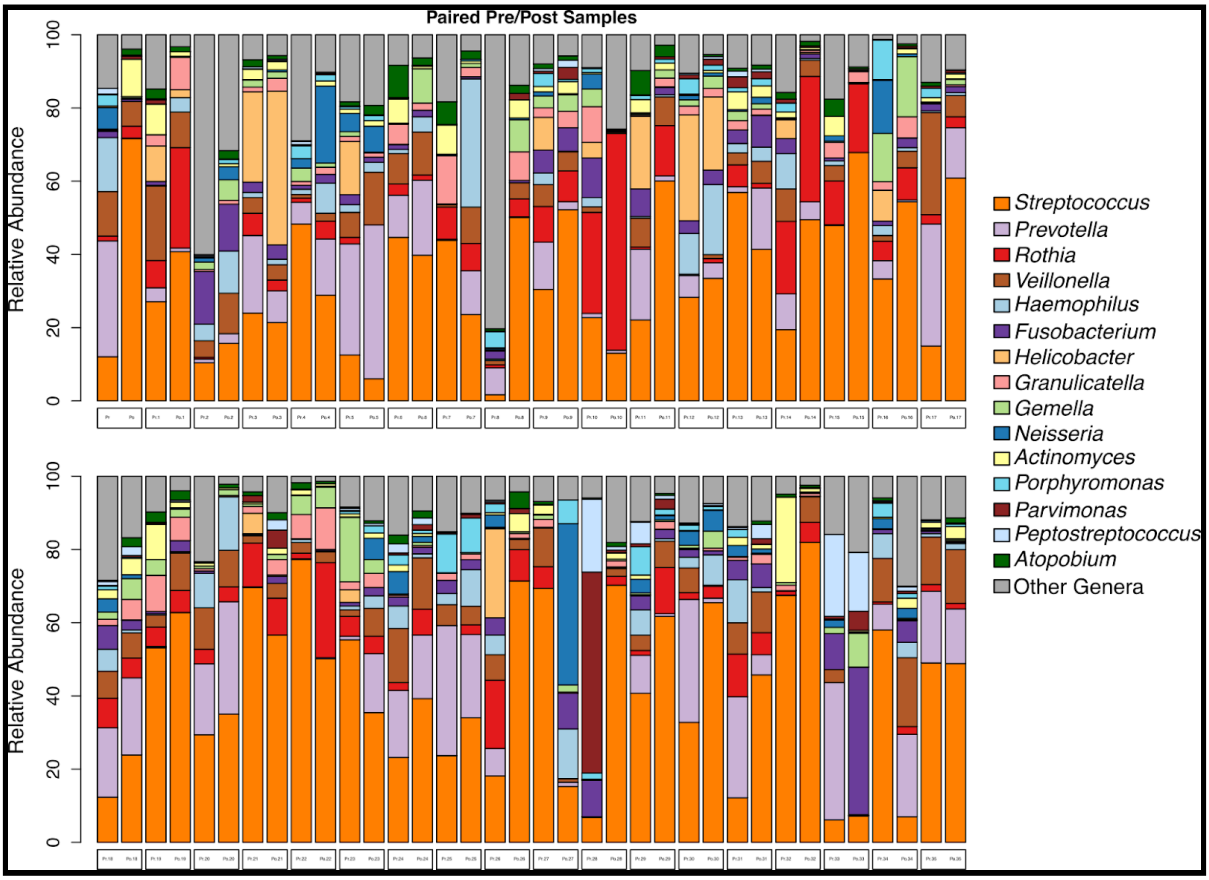


Figure 13 - Genera relative abundances before and after chemotherapy treatment. Barplots showing relative abundances for the 15 most abundant genera amongst patients. Samples from the same patient are surrounded by boxes.

The most abundant species in patients before chemotherapy treatment included *Helicobacter pylori*, *Rothia mucilaginosa* and *Prevotella melaninogenica* (which together accounted for an average of 12.9% relative abundance), whereas species most abundant in patients after chemotherapy treatment included *Rothia mucilaginosa*, *Veillonella dispar* and *Prevotella melaninogenica* (which together accounted for an average of 11.4% relative abundance). After multiple hypothesis testing correction, we found no species to be differentially abundant when comparing patient samples before and after chemotherapy treatment. Of note, HP abundance did seem to be modulated by chemotherapy treatment (p = 0.007; adjusted P-value

0.34), with 30.6% patients exhibiting decreased abundance and 2.7% increased abundance after treatment (**Figure 14)**.

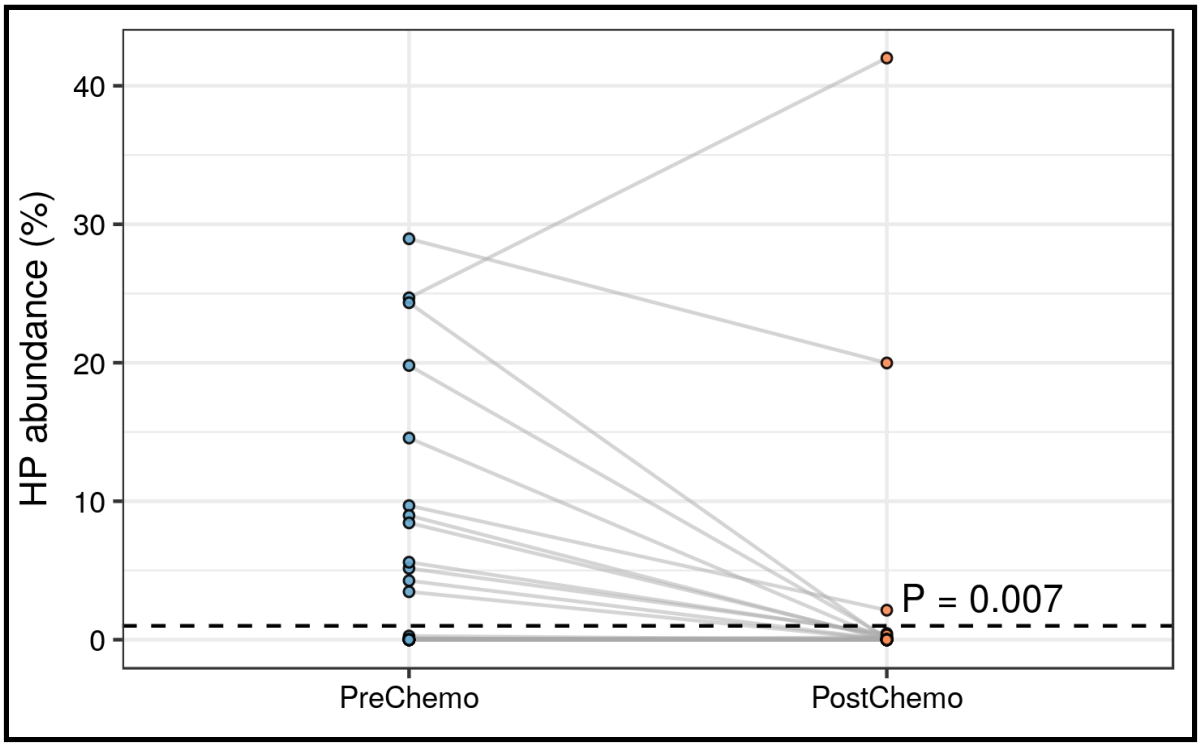


Figure 14 - Relative abundance of *Helicobacter pylori* before and after treatment. Points represent individual samples and lines connecting points represent values for the same patient before and after treatment. Dashed line indicates the relative abundance threshold of 1% for determining presence or absence of HP.

4.2.4 Contrasting neoadjuvant responders x non-responders

When assessing treatment response either via downgrading or pathological response, 7 and 5 patients were considered to respond to neoadjuvant treatment, whereas 7 and 31 patients were considered to have no response, respectively. Therefore, we investigated differences in microbial communities between these two sets of patients before treatment.

We found no differences in terms of alpha diversity between responders and non responders before treatment ($p > 0.05$). Differences in alpha diversity before and after treatment were heterogeneous between responders and non responders, with 60% of both responders and non responders exhibiting decreases in alpha diversity after treatment (**Figure 15**). However, evenness did seem to be decreased in responders compared to non responders.

When analyzing genera level abundances, two genera, *Veillonella* and *Oribacterium*, presented significant differences between the groups ($p < 0.05$), but were not significant after multiple hypothesis testing correction (**Figure 15**). When analyzing species level abundances, only one species was differentially abundant between the groups, *Oribacterium sinus*, but again, after multiple hypothesis testing correction this difference was no longer significant.

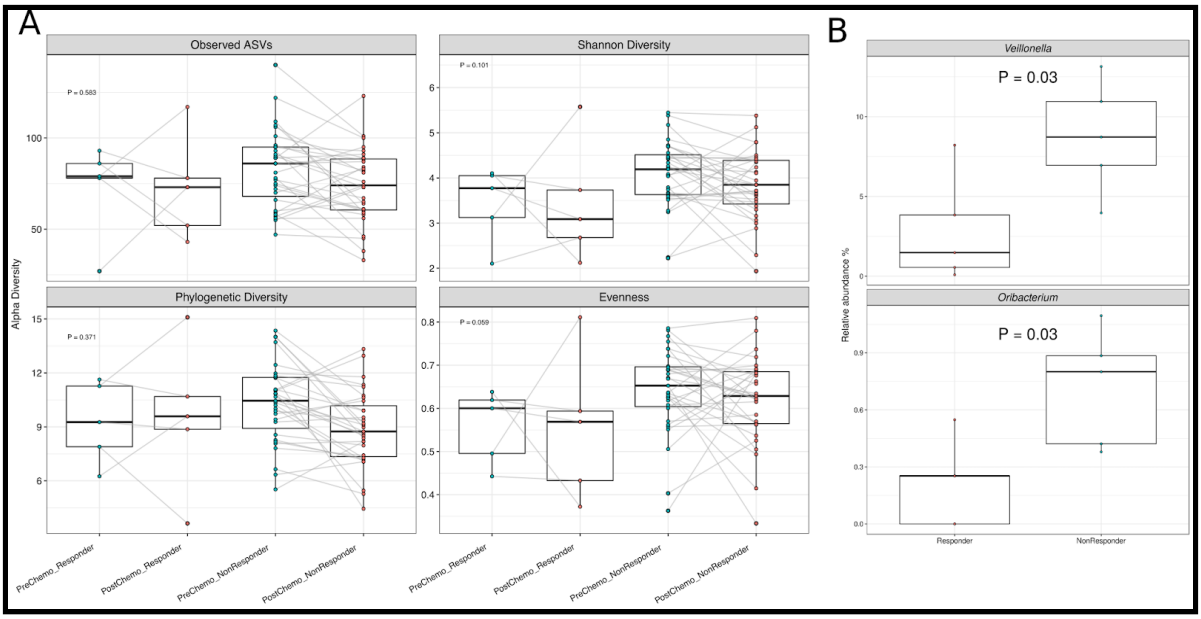


Figure 15 - Alpha diversity and genera abundances in treatment responders and non responders. (A) Boxplots showing alpha diversity in gastric adenocarcinoma samples before and after treatment separated according to treatment response using different metrics (Observed ASVs, Shannon index, Phylogenetic Diversity and Evenness - Pielou's E). Points represent individual samples and lines connecting points represent values for the same patient before and after treatment. P-values were calculated by wilcoxon rank sum tests between responders and non responders prior to treatment. **(B)** Boxplots showing relative abundances of *Veillonella* and *Oribacterium* amongst treatment responders and non responders before treatment.

4.3 Discussion

Microbial communities of gastric juice/wash samples from gastric adenocarcinoma samples presented remarkably high inter-individual variation, with relatively lower intra-individual variation, as seen by genera level abundances and distances between samples. This variation could be caused by pH, which was significantly associated with microbial communities in all distance metrics, as well as other factors such as HP presence, diet, age and BMI. Recent studies have shown that HP infection leads to significant changes in microbial diversity and dominates the gastric microbiota (Klymiuk et al. 2017; Ferreira et al. 2018). In our study, the presence of HP did not have a significant effect on gastric wash microbial communities. This could be explained by the low sample size of HP positive samples as well as the sampling method, since most studies used biopsy tissue samples.

Changes in alpha diversity due to treatment were heterogeneous amongst patients, further highlighting the inter-individual variability, with patients presenting significant decreases in richness and phylogenetic diversity after treatment. Reduced diversity due to chemotherapy treatment using fecal samples collected form non-Hodgkin's lymphoma patients has been reported previously, leading to infectious complications such as bloodstream infections and diarrhea (Montassier et al. 2015; Montassier et al. 2014) and may be a common occurrence during treatment.

We found *Streptococcus* to be the most dominant genus amongst samples, in concordance with previous studies investigating gastric microbial communities of HP

negative samples (Klymiuk et al. 2017; Ferreira et al. 2018; Coker et al. 2018). The most abundant species detected in the samples are considered to be oral microbes, with an enrichment of oral taxa being reported in gastric cancer samples compared to atrophic gastritis and intestinal metaplasia (Coker et al. 2018). This indicates that the presence and abundance of oral taxa in gastric cancer might be important for tumor initiation and progression. At various taxonomic levels we found no differences in abundance before and after treatment, possibly due to the small effect size of treatment on microbial communities and the large inter-individual variability. Despite these limitations, we observed decreases in HP abundance after chemotherapy treatment in HP positive patients. HP-positive gastric cancer patients have been shown to have a more favourable outcome when compared HP-negative patients, possibly implying that the host immune system is modulated by HP enhancing chemotherapeutic efficacy (Nishizuka et al. 2018). Out of the 14 patients we were able to categorize into treatment responders and non responders using downgrading evaluation, 3 patients who responded were HP positive whereas 2 non responders were HP positive.

When we evaluated differences in microbial communities between responders and non responders to neoadjuvant chemotherapy, we found no significant differences in alpha diversity. We observed differences in *Veillonella* and *Oribacterium* abundances between the groups, which after multiple hypothesis testing correction were non significant. The limited sample size used in this analysis may have hindered the Wilcoxon test's sensitivity to detect differences between the groups (Weiss et al. 2017; Hawinkel et al. 2017), however some interesting trends emerged, warranting further analysis and increased sample sizes.

Publication status. The results presented in this chapter will be part of a larger article where we will explore the microbial communities in gastric cancer from different perspectives, which include comparing pre-cancerous lesions with gastric cancer and different histological subtypes. As of November 20th, 2018, this study is still ongoing.

5. Conclusions

In this thesis we investigated possible roles of the microbiome in different cancer related scenarios, providing hypotheses and insights of direct or indirect effects the microbiome may have in tumor initiation, promotion and response to therapy.

Despite limitations inherent to differing experimental choices (fecal *versus* tissue samples and 16S rRNA amplicon sequencing *versus* WMS sequencing) and tumor biology (differences between proximal, distal and rectal tumors), common aspects between the studies emerged. We found increased bacterial richness in rectal tumors compared to normal tissue biopsies derived from control patients (Chapter 2), which was also seen using a different approach and a larger sample size in the meta-analysis of fecal CRC metagenomes (Chapter 3). This increased richness was partially due to the influx of species of putative oral origin, warranting further investigations to see whether these species are indeed found in the oral cavity of CRC patients and whether they are the same strain. Further investigations can also answer the timing these oral species reach the lower GI tract (pre-tumor development or post) and the route (blood or upper GI tract). Species and genera such as *Bacteroides fragilis*, *Fusobacterium, Bilophila* and *Desulfovibrio*, found to be more abundant in rectal tumors (Chapter 2), were concordantly more abundant in the meta-analysis of fecal CRC metagenomes and provide further support of their role in CRC, as biomarkers of disease presence or drivers of carcinogenesis.

The use of fecal WMS sequencing provides certain advantages over 16S rRNA sequencing, which now seem necessary to advance current knowledge of the role of microbiome in CRC. An example of such an advantage is the ability to evaluate the functional potential of the microbiome, which was crucial in determining that the abundance of the choline TMA-lyase *cutC* was increased in CRC, a novel finding, and that the most prevalent *cutC* variant found in control samples came from an unknown uncultured *Eubacterium* spp (Chapter 3). This result, along with other recent evidence (Wirbel et al. 2019), points to bacterial metabolism as a strong candidate for tumor initiation and promotion, warranting future studies to investigate the association of bacterial metabolites and transcripts with CRC, possibly via the metatranscriptomics and metametabolomics. Another advantage of using fecal WMS sequencing is the possibility to profile microbes from known species with strain-level resolution, which enabled us to evaluate possible associations between strain population structures and CRC. Using SNVs we found strong associations between strain population structures and geographical location, but no apparent disease-associated strains. When using a genomic composition-based approach, we found that CRC samples tended to possess a dominant *Bacteroides fragilis* species with a different genomic composition than that of controls. The analysis of strain

population structures in the gut microbiome is still limited, as current sequencing depths cannot provide sufficient coverage of low abundant species and can only provide information on the dominant strain, but they nevertheless provide a path to deepen our understanding of the relationship between the gut microbiome and CRC.

We also investigated whether microbial communities present in the gastric wash of GC patients were modulated by neoadjuvant chemotherapy treatment or were different in treatment responders and non responders (Chapter 4). The microbiome's role in the response to immunotherapy has gained significant attention lately, as evidence suggests it to be an important factor when assessing treatment outcome. Our results indicate a high variability between patients, possibly higher than that observed in the gut microbiome, affecting the consistency and reproducibility of our findings. Despite this limitation, preliminary analysis indicated a high prevalence and abundance of bacteria of putative oral origin in the samples; whether these bacteria are in fact active members of the community remains to be investigated.

A common theme emerged in this work, which is the presence of bacteria of putative oral origin in the different cancer types. Evidence supporting the role of the human oral microbiome in both oral cavity and whole-body diseases has been accumulating (Gao et al. 2018), with associations of oral dysbiosis in gastrointestinal diseases such as irritable bowel syndrome (IBD) (Atarashi et al. 2017), immune diseases such as rheumatoid arthritis (Zhang et al. 2015) and endocrine diseases such as type II diabetes (Xiao et al. 2017). Therefore, the oral cavity may serve as a reservoir for potential gastro-intestinal pathobionts that, when translocated to other body sites, can exacerbate both intestinal and systemic diseases. This warrants future studies to include the sampling of the oral cavity when investigating microbiome/disease associations to assess whether oral dysbiosis may also be a driving factor in disease.

The effects of non-antibiotic drugs on the gut microbiome have been largely overlooked. A recent study investigated the effects of more than 1,000 marketed drugs against 40 representative gut bacterial strains, finding that metformin, commonly used for the treatment of type II diabetes, proton pump inhibitors (PPIs) and antipsychotics inhibited the growth of at least one strain *in vitro* (Maier et al. 2018). PPIs, which are drugs that reduce acid secretion in the stomach and are therefore commonly used to treat gastroesophageal reflux, have also been shown to affect the gut microbiome, with an over-representation of multiple oral bacteria in the faecal microbiome of PPI-users (Imhann et al. 2016). As PPIs lead to increased gastric pH, oral bacterial translocation will certainly be facilitated in this scenario and could be one of the ways these bacteria reach the stomach and/or gut. Therefore, future human microbiome studies will need to carefully catalogue and account for the use of non-antibiotic drug use in their samples.

5.1 Scientific contributions

Below are the contributions to scientific knowledge as a direct (marked with an asterisk) or indirect consequence of the work developed during this thesis:

1. Kinker GS, Thomas AM, Carvalho VJ, Lima FP, Fujita A. (2016). **Deletion and low expression of NFKBIA are associated with poor prognosis in lower-grade glioma patients**. *Scientific Reports* 6:24160.
2. *Thomas AM, Jesus EC, Lopes A, Aguiar S Jr., Begnami MD, Rocha RM, Carpinetti PA, Camargo AA, Hoffmann C, Freitas HC, Silva IT, Nunes DN, Setubal JC, Dias-Neto E. (2016). **Tissue-Associated Bacterial Alterations in Rectal Carcinoma Patients Revealed by 16S rRNA Community Profiling**. *Front. Cell. Infect. Microbiol.* 6:179.
3. Antunes LP, Martins LF, Pereira RV, Thomas AM, Barbosa D, Lemos LN, Silva GM, Moura LM, Epamino GW, Digiampietri LA, Lombardi KC, Ramos PL, Quaggio RB, de Oliveira JC, Pascon RC, Cruz JB, da Silva AM, Setubal JC. (2016). **Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics**. *Scientific Reports* 6:38915.
4. *GE4GAC group. (2017). **Genomics and epidemiology for gastric adenocarcinomas**. *Applied Cancer Research* 37:7.
5. Mendes E, Acetturi BG, Thomas AM, Martins FS, Crisma AR, Murata G, Braga TT, Camâra NOS, Franco ALS, Setubal JC, Ribeiro WR, Valduga CJ, Curi R, Dias-Neto E, Tavares-de-Lima W, Ferreira CM. (2017). **Prophylactic Supplementation of Bifidobacterium longum 51A Protects Mice from Ovariectomy-Induced Exacerbated Allergic Airway Inflammation and Airway Hyperresponsiveness**. *Front. Microbiol*. 8:1732.
6. Thomas AM, Lima FP, Moura LMS, Silva AM, Dias-Neto E, Setubal JC. (2018). **Comparative Metagenomics**. *Methods Mol Biol* 1704:243-260.
7. *Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, Segata N. (2019). **Combined metagenomic analysis of colorectal cancer datasets defines cross-cohort microbial diagnostic signatures and a link with choline degradation**. *Nat med*. *Accepted*.

6. References

Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, et al. 2012. “Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome.” *PLoS Computational Biology* 8 (6): e1002358.

Ahn, Jiyoung, Rashmi Sinha, Zhiheng Pei, Christine Dominianni, Jing Wu, Jianxin Shi, James J. Goedert, Richard B. Hayes, and Liying Yang. 2013. “Human Gut Microbiome and Risk for Colorectal Cancer.” *Journal of the National Cancer Institute* 105 (24): 1907–11.

Allard, G., Ryan FJ., Jeffery I.B., Claesson M.J. 2015. “SPINGO: a rapid species-classifier for microbial amplicon sequences.” *BMC Bioinformatics* 16 (324).

Alexander, James L., Ian D. Wilson, Julian Teare, Julian R. Marchesi, Jeremy K. Nicholson, and James M. Kinross. 2017. “Gut Microbiota Modulation of Chemotherapy Efficacy and Toxicity.” *Nature Reviews. Gastroenterology & Hepatology* 14 (6): 356–65.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Amir, Amnon, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, et al. 2017. “Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.” *mSystems* 2 (2). https://doi.org/10.1128/mSystems.00191-16.

Andres-Franch, Maria, Antonio Galiana, Victoria Sanchez-Hellin, Enrique Ochoa, Eva Hernandez-Illan, Pilar Lopez-Garcia, Adela Castillejo, et al. 2017. “Streptococcus Gallolyticus Infection in Colorectal Cancer and Association with Biological and Clinical Factors.” *PloS One* 12 (3): e0174305.

Apweiler, Rolf, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, et al. 2004. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 32 (Database issue): D115–19.

Araújo-Pérez, Félix, Amber N. McCoy, Charles Okechukwu, Ian M. Carroll, Kevin M. Smith, Kim Jeremiah, Robert S. Sandler, Gary N. Asher, and Temitope O. Keku. 2012. “Differences in Microbial Signatures between Rectal Mucosal Biopsies and Rectal Swabs.” *Gut Microbes* 3 (6): 530–35.

Asnicar, Francesco, George Weingart, Timothy L. Tickle, Curtis Huttenhower, and Nicola Segata. 2015. “Compact Graphical Representation of Phylogenetic Data and Metadata with GraPhlAn.” *PeerJ* 3 (June): e1029.

Atarashi, Koji, Wataru Suda, Chengwei Luo, Takaaki Kawaguchi, Iori Motoo, Seiko Narushima, Yuya Kiguchi, et al. 2017. “Ectopic Colonization of Oral Bacteria in the Intestine Drives TH1 Cell Induction and Inflammation.” *Science* 358 (6361): 359–65.

Bae, Sajin, Cornelia M. Ulrich, Marian L. Neuhouser, Olga Malysheva, Lynn B. Bailey, Liren Xiao, Elissa C. Brown, et al. 2014. “Plasma Choline Metabolites and Colorectal Cancer Risk in the Women’s Health Initiative Observational Study.” *Cancer Research* 74 (24): 7442–52.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Binefa, Gemma, Francisco Rodríguez-Moranta, Alex Teule, and Manuel Medina-Hayas.

2014. “Colorectal Cancer: From Prevention to Personalized Medicine.” *World Journal of Gastroenterology: WJG* 20 (22): 6786–6808.
Boleij, Annemarie, Marleen M. H. J. van Gelder, Dorine W. Swinkels, and Harold Tjalsma. 2011. “Clinical Importance of Streptococcus Gallolyticus Infection among Colorectal Cancer Patients: Systematic Review and Meta-Analysis.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 53 (9): 870–78.
Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2018. “QIIME 2: Reproducible, Interactive, Scalable, and Extensible Microbiome Data Science.” e27295v1. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27295v1.
Bonder, Marc Jan, Alexander Kurilshikov, Ettje F. Tigchelaar, Zlatan Mujagic, Floris Imhann, Arnau Vich Vila, Patrick Deelen, et al. 2016. “The Effect of Host Genetics on the Gut Microbiome.” *Nature Genetics* 48 (October): 1407.
Bonnet, Régis, Antonia Suau, Joël Doré, Glenn R. Gibson, and Matthew D. Collins. 2002. “Differences in rDNA Libraries of Faecal Bacteria Derived from 10- and 25-Cycle PCRs.” *International Journal of Systematic and Evolutionary Microbiology* 52 (Pt 3): 757–63.
Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.
Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, et al. 2016. “Mobile Genes in the Human Microbiome Are Structured from Global to Individual Scales.” *Nature* 535 (7612): 435–39.
Bultman, Scott J., and Christian Jobin. 2014. “Microbial-Derived Butyrate: An Oncometabolite or Tumor-Suppressive Metabolite?” *Cell Host & Microbe* 16 (2): 143–45.
Caporaso, J. Gregory, Kyle Bittinger, Frederic D. Bushman, Todd Z. DeSantis, Gary L. Andersen, and Rob Knight. 2010. “PyNAST: A Flexible Tool for Aligning Sequences to a Template Alignment.” *Bioinformatics* 26 (2): 266–67.
Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods* 7 (5): 335–36.
Cenitagoya, G. F., C. K. Bergh, and J. Klinger-Roitman. 1998. “A Prospective Study of Gastric Cancer. ‘Real’ 5-Year Survival Rates and Mortality Rates in a Country with High Incidence.” *Digestive Surgery* 15 (4): 317–22.
Chung, Liam, Erik Thiele Orberg, Abby L. Geis, June L. Chan, Kai Fu, Christina E. DeStefano Shields, Christine M. Dejea, et al. 2018. “Bacteroides Fragilis Toxin Coordinates a Pro-Carcinogenic Inflammatory Cascade via Targeting of Colonic Epithelial Cells.” *Cell Host & Microbe* 23 (2): 203–14.e5.
Coker, Olabisi Oluwabukola, Zhenwei Dai, Yongzhan Nie, Guijun Zhao, Lei Cao, Geicho Nakatsu, William Kk Wu, et al. 2018. “Mucosal Microbiome Dysbiosis in Gastric Carcinogenesis.” *Gut* 67 (6): 1024–32.
Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. “Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis.” *Nucleic Acids Research* 42 (Database issue): D633–42.
Costea, Paul I., Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, et al. 2017. “Towards Standards for Human Fecal Sample Processing in Metagenomic Studies.” *Nature Biotechnology* 35 (October): 1069.
Cougnoux, Antony, Guillaume Dalmasso, Ruben Martinez, Emmanuel Buc, Julien Delmas,

Lucie Gibold, Pierre Sauvanet, et al. 2014. “Bacterial Genotoxin Colibactin Promotes Colon Tumour Growth by Inducing a Senescence-Associated Secretory Phenotype.” *Gut* 63 (12): 1932–42.

Craciun, Smaranda, and Emily P. Balskus. 2012. “Microbial Conversion of Choline to Trimethylamine Requires a Glycyl Radical Enzyme.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (52): 21307–12.

Craciun, Smaranda, Jonathan A. Marks, and Emily P. Balskus. 2014. “Characterization of Choline Trimethylamine-Lyase Expands the Chemistry of Glycyl Radical Enzymes.” *ACS Chemical Biology* 9 (7): 1408–13.

Crew, Katherine D., and Alfred I. Neugut. 2006. “Epidemiology of Gastric Cancer.” *World Journal of Gastroenterology: WJG* 12 (3): 354–62.

Cuevas-Ramos, Gabriel, Claude R. Petit, Ingrid Marcq, Michèle Boury, Eric Oswald, and Jean-Philippe Nougayrède. 2010. “Escherichia Coli Induces DNA Damage in Vivo and Triggers Genomic Instability in Mammalian Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (25): 11537–42.

Dai, Zhenwei, Olabisi Oluwabukola Coker, Geicho Nakatsu, William K. K. Wu, Liuyang Zhao, Zigui Chen, Francis K. L. Chan, et al. 2018. “Multi-Cohort Analysis of Colorectal Cancer Metagenome Identified Altered Bacteria across Populations and Universal Bacterial Markers.” *Microbiome* 6 (1): 70.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58.

David, Lawrence A., Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, et al. 2014. “Diet Rapidly and Reproducibly Alters the Human Gut Microbiome.” *Nature* 505 (7484): 559–63.

Dejea, Christine M., Payam Fathi, John M. Craig, Annemarie Boleij, Rahwa Taddese, Abby L. Geis, Xinqun Wu, et al. 2018. “Patients with Familial Adenomatous Polyposis Harbor Colonic Biofilms Containing Tumorigenic Bacteria.” *Science* 359 (6375): 592–97.

DerSimonian, R., and N. Laird. 1986. “Meta-Analysis in Clinical Trials.” *Controlled Clinical Trials* 7 (3): 177–88.

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): 5069–72.

Di Martino, Maria Letizia, Rosaria Campilongo, Mariassunta Casalino, Gioacchino Micheli, Bianca Colonna, and Gianni Prosseda. 2013. “Polyamines: Emerging Players in Bacteria-Host Interactions.” *International Journal of Medical Microbiology: IJMM* 303 (8): 484–91.

Drewes, Julia L., James R. White, Christine M. Dejea, Payam Fathi, Thevambiga Iyadorai, Jamuna Vadivelu, April C. Roslani, et al. 2017. “High-Resolution Bacterial 16S rRNA Gene Profile Meta-Analysis and Biofilm Status Reveal Common Colorectal Cancer Consortia.” *NPJ Biofilms and Microbiomes* 3 (November): 34.

Durbán, Ana, Juan J. Abellán, Nuria Jiménez-Hernández, Marta Ponce, Julio Ponce, Teresa Sala, Giuseppe D’Auria, Amparo Latorre, and Andrés Moya. 2011. “Assessing Gut Microbial Diversity from Feces and Rectal Mucosa.” *Microbial Ecology* 61 (1): 123–33.

Dzutsev, Amiran, Romina S. Goldszmid, Sophie Viaud, Laurence Zitvogel, and Giorgio Trinchieri. 2015. “The Role of the Microbiota in Inflammation, Carcinogenesis, and Cancer Therapy.” *European Journal of Immunology* 45 (1): 17–31.

Earley, Helen, Grainne Lennon, Aine Balfe, Michelle Kilcoyne, Marguerite Clyne, Lokesh Joshi, Stephen Carrington, et al. 2015. “A Preliminary Study Examining the Binding Capacity of Akkermansia Muciniphila and Desulfovibrio Spp., to Colonic Mucin in Health and Ulcerative Colitis.” *PloS One* 10 (10): e0135280.

Edgar, Robert C. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics* 26 (19): 2460–61.

Edgar, Robert C. 2013. “UPARSE: Highly Accurate OTU Sequences from Microbial Amplicon Reads.” *Nature Methods* 10 (August): 996.

Edgar, Robert C., Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight. 2011. “UCHIME Improves Sensitivity and Speed of Chimera Detection.” *Bioinformatics* 27 (16): 2194–2200.

Faith, Daniel P., and Andrew M. Baker. 2007. “Phylogenetic Diversity (PD) and Biodiversity Conservation: Some Bioinformatics Challenges.” *Evolutionary Bioinformatics Online* 2 (February): 121–28.

Faith, Jeremiah J., Janaki L. Guruge, Mark Charbonneau, Sathish Subramanian, Henning Seedorf, Andrew L. Goodman, Jose C. Clemente, et al. 2013. “The Long-Term Stability of the Human Gut Microbiota.” *Science* 341 (6141): 1237439.

Feng, Qiang, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, et al. 2015. “Gut Microbiome Development along the Colorectal Adenoma–carcinoma Sequence.” *Nature Communications* 6 (March): 6528.

Ferreira, Rui M., Joana Pereira-Marques, Ines Pinto-Ribeiro, Jose L. Costa, Fatima Carneiro, Jose C. Machado, and Ceu Figueiredo. 2018. “Gastric Microbial Community Profiling Reveals a Dysbiotic Cancer-Associated Microbiota.” *Gut* 67 (2): 226–36.

Fijan, Sabina. 2014. “Microorganisms with Claimed Probiotic Properties: An Overview of Recent Literature.” *International Journal of Environmental Research and Public Health* 11 (5): 4745–67.

Flemer, Burkhardt, Denise B. Lynch, Jillian M. R. Brown, Ian B. Jeffery, Feargal J. Ryan, Marcus J. Claesson, Micheal O’Riordain, Fergus Shanahan, and Paul W. O’Toole. 2017. “Tumour-Associated and Non-Tumour-Associated Microbiota in Colorectal Cancer.” *Gut* 66 (4): 633–43.

Flemer, Burkhardt, Ryan D. Warren, Maurice P. Barrett, Katryna Cisek, Anubhav Das, Ian B. Jeffery, Eimear Hurley, Micheal O’Riordain, Fergus Shanahan, and Paul W. O’Toole. 2017. “The Oral Microbiota in Colorectal Cancer Is Distinctive and Predictive.” *Gut*, October. https://doi.org/10.1136/gutjnl-2017-314814.

Foulkes, William D. 2008. “Inherited Susceptibility to Common Cancers.” *The New England Journal of Medicine* 359 (20): 2143–53.

Frank, Christoph, Jan Sundquist, Hongyao Yu, Akseli Hemminki, and Kari Hemminki. 2017. “Concordant and Discordant Familial Cancer: Familial Risks, Proportions and Population Impact.” *International Journal of Cancer. Journal International Du Cancer* 140 (7): 1510–16.

Fuerst, John A., and Evgeny Sagulenko. 2011. “Beyond the Bacterium: Planctomycetes Challenge Our Concepts of Microbial Structure and Function.” *Nature Reviews. Microbiology* 9 (6): 403–13.

Fulbright, Laura E., Melissa Ellermann, and Janelle C. Arthur. 2017. “The Microbiome and the Hallmarks of Cancer.” *PLoS Pathogens* 13 (9): e1006480.

Furet, Jean-Pierre, Olivier Firmesse, Michèle Gourmelon, Chantal Bridonneau, Julien Tap, Stanislas Mondot, Joël Doré, and Gérard Corthier. 2009. “Comparative Assessment of Human and Farm Animal Faecal Microbiota Using Real-Time Quantitative PCR.”

FEMS Microbiology Ecology 68 (3): 351–62.
Galluzzi, Lorenzo, Aitziber Buqué, Oliver Kepp, Laurence Zitvogel, and Guido Kroemer. 2015. “Immunological Effects of Conventional Chemotherapy and Targeted Anticancer Agents.” *Cancer Cell* 28 (6): 690–714.
Galvan, Antonella, John P. A. Ioannidis, and Tommaso A. Dragani. 2010. “Beyond Genome-Wide Association Studies: Genetic Heterogeneity and Individual Predisposition to Cancer.” *Trends in Genetics: TIG* 26 (3): 132–41.
Gao, Lu, Tiansong Xu, Gang Huang, Song Jiang, Yan Gu, and Feng Chen. 2018. “Oral Microbiomes: More and More Importance in Oral Cavity and Whole Body.” *Protein & Cell*, May. https://doi.org/10.1007/s13238-018-0548-1.
Gao, Zhiguang, Bomin Guo, Renyuan Gao, Qingchao Zhu, and Huanlong Qin. 2015. “Microbiota Disbiosis Is Associated with Colorectal Cancer.” *Frontiers in Microbiology* 6 (February): 20.
Geller, Leore T., Michal Barzily-Rokni, Tal Danino, Oliver H. Jonas, Noam Shental, Deborah Nejman, Nancy Gavert, et al. 2017. “Potential Role of Intratumor Bacteria in Mediating Tumor Resistance to the Chemotherapeutic Drug Gemcitabine.” *Science* 357 (6356): 1156–60.
Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. “The Treatment-Naive Microbiome in New-Onset Crohn’s Disease.” *Cell Host & Microbe* 15 (3): 382–92.
Gill, Steven R., Derrick E. Fouts, Gordon L. Archer, Emmanuel F. Mongodin, Robert T. Deboy, Jacques Ravel, Ian T. Paulsen, et al. 2005. “Insights on Evolution of Virulence and Resistance from the Complete Genome Analysis of an Early Methicillin-Resistant Staphylococcus Aureus Strain and a Biofilm-Producing Methicillin-Resistant Staphylococcus Epidermidis Strain.” *Journal of Bacteriology* 187 (7): 2426–38.
Greenblum, Sharon, Rogan Carr, and Elhanan Borenstein. 2015. “Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species.” *Cell* 160 (4): 583–94.
Grivennikov, Sergei I., Kepeng Wang, Daniel Mucida, C. Andrew Stewart, Bernd Schnabl, Dominik Jauch, Koji Taniguchi, et al. 2012. “Adenoma-Linked Barrier Defects and Microbial Products Drive IL-23/IL-17-Mediated Tumour Growth.” *Nature* 491 (7423): 254–58.
Guertin, Kristin A., Xinmin S. Li, Barry I. Graubard, Demetrius Albanes, Stephanie J. Weinstein, James J. Goedert, Zeneng Wang, Stanley L. Hazen, and Rashmi Sinha. 2017. “Serum Trimethylamine N-Oxide, Carnitine, Choline, and Betaine in Relation to Colorectal Cancer Risk in the Alpha Tocopherol, Beta Carotene Cancer Prevention Study.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 26 (6): 945–52.
Gur, Chamutal, Yara Ibrahim, Batya Isaacson, Rachel Yamin, Jawad Abed, Moriya Gamliel, Jonatan Enk, et al. 2015. “Binding of the Fap2 Protein of Fusobacterium Nucleatum to Human Inhibitory Receptor TIGIT Protects Tumors from Immune Cell Attack.” *Immunity* 42 (2): 344–55.
Hajishengallis, George, Richard P. Darveau, and Michael A. Curtis. 2012. “The Keystone-Pathogen Hypothesis.” *Nature Reviews. Microbiology* 10 (10): 717–25.
Hannigan, Geoffrey D., Melissa B. Duhaime, Mack T. Ruffin, Charlie C. Koumpouras, and Patrick D. Schloss. 2017. “Viral and Bacterial Communities of Colorectal Cancer.” *bioRxiv*. https://doi.org/10.1101/152868.
Hardbower, Dana M., Thibaut de Sablet, Rupesh Chaturvedi, and Keith T. Wilson. 2013.

"Chronic Inflammation and Oxidative Stress: The Smoking Gun for Helicobacter Pylori-Induced Gastric Cancer?" *Gut Microbes* 4 (6): 475–81.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Vol. 1. 0172-7397. New York: Springer-Verlag New York.
Hawinkel, Stijn, Federico Mattiello, Luc Bijnens, and Olivier Thas. 2017. "A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery Rate." *Briefings in Bioinformatics*, August. https://doi.org/10.1093/bib/bbx104.
Holmes, Ian, Keith Harris, and Christopher Quince. 2012. "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics." *PloS One* 7 (2): e30126.
Hong, Pei-Ying, Jennifer A. Croix, Eugene Greenberg, H. Rex Gaskins, and Roderick I. Mackie. 2011. "Pyrosequencing-Based Analysis of the Mucosal Microbiota in Healthy Individuals Reveals Ubiquitous Bacterial Groups and Micro-Heterogeneity." *PloS One* 6 (9): e25042.
Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207–14.
Huxley, Rachel R., Alireza Ansary-Moghaddam, Peter Clifton, Sebastien Czernichow, Christine L. Parr, and Mark Woodward. 2009. "The Impact of Dietary and Lifestyle Risk Factors on Risk of Colorectal Cancer: A Quantitative Overview of the Epidemiological Evidence." *International Journal of Cancer. Journal International Du Cancer* 125 (1): 171–80.
Imhann, Floris, Marc Jan Bonder, Arnau Vich Vila, Jingyuan Fu, Zlatan Mujagic, Lisa Vork, Ettje F. Tigchelaar, et al. 2016. "Proton Pump Inhibitors Affect the Gut Microbiome." *Gut* 65 (5): 740–48.
Jandhyala, Sai Manasa, Rupjyoti Talukdar, Chivkula Subramanyam, Harish Vuyyuru, Mitnala Sasikala, and D. Nageshwar Reddy. 2015. "Role of the Normal Gut Microbiota." *World Journal of Gastroenterology: WJG* 21 (29): 8787–8803.
Jie, Zhuye, Huihua Xia, Shi-Long Zhong, Qiang Feng, Shenghui Li, Suisha Liang, Huanzi Zhong, et al. 2017. "The Gut Microbiome in Atherosclerotic Cardiovascular Disease." *Nature Communications* 8 (1): 845.
Johnson, Constance M., Caimiao Wei, Joe E. Ensor, Derek J. Smolenski, Christopher I. Amos, Bernard Levin, and Donald A. Berry. 2013. "Meta-Analyses of Colorectal Cancer Risk Factors." *Cancer Causes & Control: CCC* 24 (6): 1207–22.
Kalnins, Gints, Janis Kuka, Solveiga Grinberga, Marina Makrecka-Kuka, Edgars Liepinsh, Maija Dambrova, and Kaspars Tars. 2015. "Structure and Function of CutC Choline Lyase from Human Microbiota Bacterium Klebsiella Pneumoniae." *The Journal of Biological Chemistry* 290 (35): 21732–40.
Kaminski, James, Molly K. Gibson, Eric A. Franzosa, Nicola Segata, Gautam Dantas, and Curtis Huttenhower. 2015. "High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED." *PLoS Computational Biology* 11 (12): e1004557.
Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.
Kelly, Brendan J., Robert Gross, Kyle Bittinger, Scott Sherrill-Mix, James D. Lewis, Ronald G. Collman, Frederic D. Bushman, and Hongzhe Li. 2015. "Power and Sample-Size Estimation for Microbiome Studies Using Pairwise Distances and PERMANOVA." *Bioinformatics* 31 (15): 2461–68.
Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias

Horn, and Frank Oliver Glöckner. 2013. “Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and next-Generation Sequencing-Based Diversity Studies.” *Nucleic Acids Research* 41 (1): e1.

Klymiuk, Ingeborg, Ceren Bilgilier, Alexander Stadlmann, Jakob Thannesberger, Marie-Theres Kastner, Christoph Högenauer, Andreas Püspök, et al. 2017. “The Human Gastric Microbiome Is Predicated upon Infection with Helicobacter Pylori.” *Frontiers in Microbiology* 8 (December): 2508.

Koeppel, Max, Fernando Garcia-Alcalde, Frithjof Glowinski, Philipp Schlaermann, and Thomas F. Meyer. 2015. “Helicobacter Pylori Infection Causes Characteristic DNA Damage Patterns in Human Cells.” *Cell Reports* 11 (11): 1703–13.

Koeth, Robert A., Bruce S. Levison, Miranda K. Culley, Jennifer A. Buffa, Zeneng Wang, Jill C. Gregory, Elin Org, et al. 2014. “γ-Butyrobetaine Is a Proatherogenic Intermediate in Gut Microbial Metabolism of L-Carnitine to TMAO.” *Cell Metabolism* 20 (5): 799–812.

Korem, Tal, David Zeevi, Jotham Suez, Adina Weinberger, Tali Avnit-Sagi, Maya Pompan-Lotan, Elad Matot, et al. 2015. “Growth Dynamics of Gut Microbiota in Health and Disease Inferred from Single Metagenomic Samples.” *Science* 349 (6252): 1101–6.

Koren, Omry, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, Rob Knight, Curtis Huttenhower, and Ruth E. Ley. 2013. “A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets.” *PLoS Computational Biology* 9 (1): e1002863.

Kostic, Aleksandar D., Eunyoung Chun, Lauren Robertson, Jonathan N. Glickman, Carey Ann Gallini, Monia Michaud, Thomas E. Clancy, et al. 2013. “Fusobacterium Nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment.” *Cell Host & Microbe* 14 (2): 207–15.

Kostic, Aleksandar D., Dirk Gevers, Chandra Sekhar Pedamallu, Monia Michaud, Fujiko Duke, Ashlee M. Earl, Akinyemi I. Ojesina, et al. 2012. “Genomic Analysis Identifies Association of Fusobacterium with Colorectal Carcinoma.” *Genome Research* 22 (2): 292–98.

Kummen, Martin, Mette Vesterhus, Marius Trøseid, Bjørn Moum, Asbjørn Svardal, Kirsten Muri Boberg, Pål Aukrust, Tom Hemming Karlsen, Rolf Kristian Berge, and Johannes Roksund Hov. 2017. “Elevated Trimethylamine-N-Oxide (TMAO) Is Associated with Poor Prognosis in Primary Sclerosing Cholangitis Patients with Normal Liver Function.” *United European Gastroenterology Journal* 5 (4): 532–41.

Lane, JD. 1991. “16S/23S rRNA Sequencing.” *Nucleic Acid Techniques in Bacterial Systematics*, 125–75.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (March): 357.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Lozupone, Catherine, and Rob Knight. 2005. “UniFrac: A New Phylogenetic Method for Comparing Microbial Communities.” *Applied and Environmental Microbiology* 71 (12): 8228–35.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2018. “Cluster: Cluster Analysis Basics and Extensions.”

Maier, Lisa, Mihaela Pruteanu, Michael Kuhn, Georg Zeller, Anja Telzerow, Exene Erin Anderson, Ana Rita Brochado, et al. 2018. “Extensive Impact of Non-Antibiotic Drugs

on Human Gut Bacteria." *Nature* 555 (7698): 623–28.
Marshall, B. J., and J. R. Warren. 1984. "Unidentified Curved Bacilli in the Stomach of Patients with Gastritis and Peptic Ulceration." *The Lancet* 1 (8390): 1311–15.
Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1): 10–12.
McGarr, Sean E., Jason M. Ridlon, and Phillip B. Hylemon. 2005. "Diet, Anaerobic Bacterial Metabolism, and Colon Cancer: A Review of the Literature." *Journal of Clinical Gastroenterology* 39 (2): 98–109.
McMurdie, Paul J., and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PloS One* 8 (4): e61217.
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." *Journal of Gastroenterology* 50 (2): 167–79.
Montassier, E., T. Gastinne, P. Vangay, G. A. Al-Ghalith, S. Bruley des Varannes, S. Massart, P. Moreau, et al. 2015. "Chemotherapy-Driven Dysbiosis in the Intestinal Microbiome." *Alimentary Pharmacology & Therapeutics* 42 (5): 515–28.
Montassier, Emmanuel, Eric Batard, Sébastien Massart, Thomas Gastinne, Thomas Carton, Jocelyne Caillon, Sophie Le Fresne, et al. 2014. "16S rRNA Gene Pyrosequencing Reveals Shift in Patient Faecal Microbiota during High-Dose Chemotherapy as Conditioning Regimen for Bone Marrow Transplantation." *Microbial Ecology* 67 (3): 690–99.
Moore, Patrick S., and Yuan Chang. 2010. "Why Do Viruses Cause Cancer? Highlights of the First Century of Human Tumour Virology." *Nature Reviews. Cancer* 10 (12): 878–89.
Moore, W. E., and L. H. Moore. 1995. "Intestinal Floras of Populations That Have a High Risk of Colon Cancer." *Applied and Environmental Microbiology* 61 (9): 3202–7.
Morgan, Xochitl C., Nicola Segata, and Curtis Huttenhower. 2013. "Biodiversity and Functional Genomics in the Human Microbiome." *Trends in Genetics: TIG* 29 (1): 51–58.
Muegge, Brian D., Justin Kuczynski, Dan Knights, Jose C. Clemente, Antonio González, Luigi Fontana, Bernard Henrissat, Rob Knight, and Jeffrey I. Gordon. 2011. "Diet Drives Convergence in Gut Microbiome Functions across Mammalian Phylogeny and within Humans." *Science* 332 (6032): 970–74.
Nakatsu, Geicho, Xiangchun Li, Haokui Zhou, Jianqiu Sheng, Sunny Hei Wong, William Ka Kai Wu, Siew Chien Ng, et al. 2015. "Gut Mucosal Microbiome across Stages of Colorectal Carcinogenesis." *Nature Communications* 6 (October): 8727.
Nelson, William C., and James C. Stegen. 2015. "The Reduced Genomes of Parcubacteria (OD1) Contain Signatures of a Symbiotic Lifestyle." *Frontiers in Microbiology* 6 (July): 713.
Nishizuka, Satoshi S., Gen Tamura, Masahiro Nakatochi, Norimasa Fukushima, Yukimi Ohmori, Chihiro Sumida, Takeshi Iwaya, Takashi Takahashi, Keisuke Koeda, and Northern Japan Gastric Cancer Study Consortium. 2018. "Helicobacter Pylori Infection Is Associated with Favorable Outcome in Advanced Gastric Cancer Patients Treated with S-1 Adjuvant Chemotherapy." *Journal of Surgical Oncology* 117 (5): 947–56.
Nobili, Stefania, Lorenzo Bruno, Ida Landini, Cristina Napoli, Paolo Bechi, Francesco Tonelli, Carlos A. Rubio, Enrico Mini, and Gabriella Nesi. 2011. "Genomic and Genetic Alterations Influence the Progression of Gastric Cancer." *World Journal of*

Gastroenterology: WJG 17 (3): 290–99.
Oellgaard, Jens, Signe Abitz Winther, Tobias Schmidt Hansen, Peter Rossing, and Bernt Johan von Scholten. 2017. “Trimethylamine N-Oxide (TMAO) as a New Potential Therapeutic Target for Insulin Resistance and Cancer.” *Current Pharmaceutical Design* 23 (25): 3699–3712.
Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2018. “Vegan: Community Ecology Package.” https://CRAN.R-project.org/package=vegan.
O’Mahony, L., M. Feeney, S. O’Halloran, L. Murphy, B. Kiely, J. Fitzgibbon, G. Lee, G. O’Sullivan, F. Shanahan, and J. K. Collins. 2001. “Probiotic Impact on Microbial Flora, Inflammation and Tumour Development in IL-10 Knockout Mice.” *Alimentary Pharmacology & Therapeutics* 15 (8): 1219–25.
Ou, Junhai, James P. DeLany, Ming Zhang, Sumit Sharma, and Stephen J. D. O’Keefe. 2012. “Association between Low Colonic Short-Chain Fatty Acids and High Bile Acids in High Colon Cancer Risk Populations.” *Nutrition and Cancer* 64 (1): 34–40.
Parkin, Donald Maxwell. 2006. “The Global Health Burden of Infection-Associated Cancers in the Year 2002.” *International Journal of Cancer. Journal International Du Cancer* 118 (12): 3030–44.
Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.” *Nature Microbiology* 2 (11): 1533–42.
Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, et al. 2017. “Accessible, Curated Metagenomic Data through ExperimentHub.” *Nature Methods* 14 (October): 1023.
Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. “Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights.” *PLoS Computational Biology* 12 (7): e1004977.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30.
Peek, Richard M., Jr, and Jean E. Crabtree. 2006. “Helicobacter Infection and Gastric Neoplasia.” *The Journal of Pathology* 208 (2): 233–48.
Peng, Hanchuan, Fuhui Long, and C. Ding. 2005. “Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8): 1226–38.
Picard, C., J. Fioramonti, A. Francois, T. Robinson, F. Neant, and C. Matuchansky. 2005. “Review Article: Bifidobacteria as Probiotic Agents -- Physiological Effects and Clinical Benefits.” *Alimentary Pharmacology & Therapeutics* 22 (6): 495–512.
Pitt, J. M., A. Marabelle, A. Eggermont, J-C Soria, G. Kroemer, and L. Zitvogel. 2016. “Targeting the Tumor Microenvironment: Removing Obstruction to Anticancer Immune Responses and Immunotherapy.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 27 (8): 1482–92.
Polk, D. Brent, and Richard M. Peek Jr. 2010. “Helicobacter Pylori: Gastric Cancer and beyond.” *Nature Reviews. Cancer* 10 (6): 403–14.
Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2009. “FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix.” *Molecular Biology and Evolution* 26 (7): 1641–50.

Price, Morgan N. 2010. “FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.
Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (March): 59.
Ramakrishna, Balakrishnan S. 2007. “The Normal Bacterial Flora of the Human Intestine and Its Regulation.” *Journal of Clinical Gastroenterology* 41: S2.
Rath, Silke, Benjamin Heidrich, Dietmar H. Pieper, and Marius Vital. 2017. “Uncovering the Trimethylamine-Producing Bacteria of the Human Gut Microbiota.” *Microbiome* 5 (1): 54.
R Core Team. 2018. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Ridlon, Jason M., Dae J. Kang, Phillip B. Hylemon, and Jasmohan S. Bajaj. 2014. “Bile Acids and the Gut Microbiome.” *Current Opinion in Gastroenterology* 30 (3): 332–38.
Riester, Markus, Wei Wei, Levi Waldron, Aedin C. Culhane, Lorenzo Trippa, Esther Oliva, Sung-Hoon Kim, et al. 2014. “Risk Prediction for Late-Stage Ovarian Cancer by Meta-Analysis of 1525 Patient Samples.” *Journal of the National Cancer Institute* 106 (5). https://doi.org/10.1093/jnci/dju048.
Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.
Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. “VSEARCH: A Versatile Open Source Tool for Metagenomics.” *PeerJ* 4 (October): e2584.
Ross, J. S., N. E. Stagliano, M. J. Donovan, R. E. Breitbart, and G. S. Ginsburg. 2001. “Atherosclerosis and Cancer: Common Molecular Pathways of Disease Development and Progression.” *Annals of the New York Academy of Sciences* 947 (December): 271–92; discussion 292–93.
Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20 (November): 53–65.
Routy, Bertrand, Emmanuelle Le Chatelier, Lisa Derosa, Connie P. M. Duong, Maryam Tidjani Alou, Romain Daillère, Aurélie Fluckiger, et al. 2017. “Gut Microbiome Influences Efficacy of PD-1–based Immunotherapy against Epithelial Tumors.” *Science*, November, eaan3706.
Rubinstein, Mara Roxana, Xiaowei Wang, Wendy Liu, Yujun Hao, Guifang Cai, and Yiping 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.
Sabino, João, Sara Vieira-Silva, Kathleen Machiels, Marie Joossens, Gwen Falony, Vera Ballet, Marc Ferrante, et al. 2016. “Primary Sclerosing Cholangitis Is Characterised by Intestinal Dysbiosis Independent from IBD.” *Gut* 65 (10): 1681–89.
Salama, N., K. Guillemin, T. K. McDaniel, G. Sherlock, L. Tompkins, and S. Falkow. 2000. “A Whole-Genome Microarray Reveals Genetic Diversity among Helicobacter Pylori Strains.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (26): 14668–73.
Sanapareddy, Nina, Ryan M. Legge, Biljana Jovov, Amber McCoy, Lauren Burcal, Felix Araujo-Perez, Thomas A. Randall, et al. 2012. “Increased Rectal Microbial Richness Is

Associated with the Presence of Colorectal Adenomas in Humans." *The ISME Journal* 6 (10): 1858–68.
Schloissnig, Siegfried, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, et al. 2012. "Genomic Variation Landscape of the Human Gut Microbiome." *Nature* 493 (December): 45.
Schloss, Patrick D., and Jo Handelsman. 2005. "Metagenomics for Studying Unculturable Microorganisms: Cutting the Gordian Knot." *Genome Biology* 6 (8): 229.
Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. 2016. "Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics." *Nature Methods* 13 (5): 435–38.
Sears, Cynthia L., Abby L. Geis, and Franck Housseau. 2014. "Bacteroides Fragilis Subverts Mucosal Biology: From Symbiont to Colon Carcinogenesis." *The Journal of Clinical Investigation* 124 (10): 4166–72.
Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–69.
Segata, Nicola, Daniela Börnigen, Xochitl C. Morgan, and Curtis Huttenhower. 2013. "PhyloPhlAn Is a New Method for Improved Phylogenetic and Taxonomic Placement of Microbes." *Nature Communications* 4: 2304.
Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, and Curtis Huttenhower. 2011. "Metagenomic Biomarker Discovery and Explanation." *Genome Biology* 12 (6): R60.
Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.
Siezen, Roland J., Vesela A. Tzeneva, Anna Castioni, Michiel Wels, Hoa T. K. Phan, Jan L. W. Rademaker, Marjo J. C. Starrenburg, Michiel Kleerebezem, Douwe Molenaar, and Johan E. T. van Hylckama Vlieg. 2010. "Phenotypic and Genomic Diversity of Lactobacillus Plantarum Strains Isolated from Various Environmental Niches." *Environmental Microbiology* 12 (3): 758–73.
Simon, Richard, Michael D. Radmacher, Kevin Dobbin, and Lisa 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.
Simpson, E. H. 1949. "Measurement of Diversity." *Nature* 163 (April): 688.
Solheim, Margrete, Agot Aakra, Lars G. Snipen, Dag A. Brede, and Ingolf F. Nes. 2009. "Comparative Genomics of Enterococcus Faecalis from Healthy Norwegian Infants." *BMC Genomics* 10 (April): 194.
Sommer, Felix, and Fredrik Bäckhed. 2013. "The Gut Microbiota--Masters of Host Development and Physiology." *Nature Reviews. Microbiology* 11 (4): 227–38.
Soo, Rochelle M., Connor T. Skennerton, Yuji Sekiguchi, Michael Imelfort, Samuel J. Paech, Paul G. Dennis, Jason A. Steen, Donovan H. Parks, Gene W. Tyson, and Philip Hugenholtz. 2014. "An Expanded Genomic Representation of the Phylum Cyanobacteria." *Genome Biology and Evolution* 6 (5): 1031–45.
Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13.
Sung, Jihee, Nayoung Kim, Jaeyeon Kim, Hyun Jin Jo, Ji Hyun Park, Ryoung Hee Nam, Yeong-Jae Seok, Yeon-Ran Kim, Dong Ho Lee, and Hyun Chae Jung. 2016. "Comparison of Gastric Microbiota Between Gastric Juice and Mucosa by Next Generation Sequencing Method." *Journal of Cancer Prevention* 21 (1): 60–65.

Tamas, K., A. M. E. Walenkamp, E. G. E. de Vries, M. A. T. M. van Vugt, R. G. Beets-Tan, B. van Etten, D. J. A. de Groot, and G. A. P. Hospers. 2015. “Rectal and Colon Cancer: Not Just a Different Anatomic Site.” *Cancer Treatment Reviews* 41 (8): 671–79.

Tanaka, S., A. Tatsuguchi, S. Futagami, K. Gudis, K. Wada, T. Seo, K. Mitsui, et al. 2006. “Monocyte Chemoattractant Protein 1 and Macrophage Cyclooxygenase 2 Expression in Colonic Adenoma.” *Gut* 55 (1): 54–61.

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

Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Pozzi, C., Gandini, S., Serrano, S., Tarallo, S., Francavilla, A., Gallo. G., Trompetto, M., Ferrero, G., Cordero, F., Dias-Neto, E., Setubal, J.C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A., Segata, N. 2019. “Combined metagenomic analysis of colorectal cancer datasets defines cross-cohort microbial diagnostic signatures and a link with choline degradation.” *Nat med*. *Accepted*.

Thomas, Ryan M., and Christian Jobin. 2015. “The Microbiome and Cancer: Is the ‘Oncobiome’ Mirage Real?” *Trends in Cancer Research* 1 (1): 24–35.

Tibshirani, Robert, and Guenther Walther. 2005. “Cluster Validation by Prediction Strength.” *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 14 (3): 511–28.

Tong, Jia, Chengxu Liu, Paula Summanen, Huaxi Xu, and Sydney M. Finegold. 2011. “Application of Quantitative Real-Time PCR for Rapid Identification of Bacteroides Fragilis Group and Related Organisms in Human Wound Samples.” *Anaerobe* 17 (2): 64–68.

Torre, Lindsey A., Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. 2015. “Global Cancer Statistics, 2012.” *CA: A Cancer Journal for Clinicians* 65 (2): 87–108.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. “MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling.” *Nature Methods* 12 (10): 902–3.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. “Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes.” *Genome Research* 27 (4): 626–38.

Turnbaugh, Peter J., Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, et al. 2008. “A Core Gut Microbiome in Obese and Lean Twins.” *Nature* 457 (November): 480.

Ulger Toprak, N., A. Yagci, B. M. Gulluoglu, M. L. Akin, P. Demirkalem, T. Celenk, and G. Soyletir. 2006. “A Possible Role of Bacteroides Fragilis Enterotoxin in the Aetiology of Colorectal Cancer.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 12 (8): 782–86.

Vétizou, Marie, Jonathan M. Pitt, Romain Daillère, Patricia Lepage, Nadine Waldschmitt, Caroline Flament, Sylvie Rusakiewicz, et al. 2015. “Anticancer Immunotherapy by CTLA-4 Blockade Relies on the Gut Microbiota.” *Science* 350 (6264): 1079–84.

Vliet, Michel J. van, Wim J. E. Tissing, Catharina A. J. Dun, Nico E. L. Meessen, Willem A.

Kamps, Eveline S. J. M. de Bont, and Hermie J. M. Harmsen. 2009. “Chemotherapy Treatment in Pediatric Patients with Acute Myeloid Leukemia Receiving Antimicrobial Prophylaxis Leads to a Relative Increase of Colonization with Potentially Pathogenic Bacteria in the Gut.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 49 (2): 262–70.

Vogtmann, Emily, Xing Hua, Georg Zeller, Shinichi Sunagawa, Anita Y. Voigt, Rajna Hercog, James J. Goedert, Jianxin Shi, Peer Bork, and Rashmi Sinha. 2016. “Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing.” *PloS One* 11 (5): e0155362.

Vollmer, Robin T. 2006. “Predictive Probability of Serum Prostate-Specific Antigen for Prostate Cancer: An Approach Using Bayes Rule.” *American Journal of Clinical Pathology* 125 (3): 336–42.

Walker, Alan W., Jennifer C. Martin, Paul Scott, Julian Parkhill, Harry J. Flint, and Karen P. Scott. 2015. “16S rRNA Gene-Based Profiling of the Human Infant Gut Microbiota Is Strongly Influenced by Sample Processing and PCR Primer Choice.” *Microbiome* 3 (June): 26.

Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole. 2007. “Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.” *Applied and Environmental Microbiology* 73 (16): 5261–67.

Wang, Zi-Kai, and Yun-Sheng Yang. 2013. “Upper Gastrointestinal Microbiota and Digestive Diseases.” *World Journal of Gastroenterology: WJG* 19 (10): 1541–50.

Ward, Tonya, Jake Larson, Jeremy Meulemans, Ben Hillmann, Joshua Lynch, Dimitri Sidiropoulos, John Spear, et al. 2017. “BugBase Predicts Organism Level Microbiome Phenotypes.” *bioRxiv*. https://doi.org/10.1101/133462.

Warren, René L., Douglas J. Freeman, Stephen Pleasance, Peter Watson, Richard A. Moore, Kyla Cochrane, Emma Allen-Vercoe, and Robert A. Holt. 2013. “Co-Occurrence of Anaerobic Bacteria in Colorectal Carcinomas.” *Microbiome* 1 (1): 16.

Wei, Esther K., Edward Giovannucci, Kana Wu, Bernard Rosner, Charles S. Fuchs, Walter C. Willett, and Graham A. Colditz. 2004. “Comparison of Risk Factors for Colon and Rectal Cancer.” *International Journal of Cancer. Journal International Du Cancer* 108 (3): 433–42.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, et al. 2017. “Normalization and Microbial Differential Abundance Strategies Depend upon Data Characteristics.” *Microbiome* 5 (1): 27.

Wick, Elizabeth C., Shervin Rabizadeh, Emilia Albesiano, Xinqun Wu, Shaoguang Wu, June Chan, Ki-Jong Rhee, et al. 2014. “Stat3 Activation in Murine Colitis Induced by Enterotoxigenic Bacteroides Fragilis.” *Inflammatory Bowel Diseases* 20 (5): 821–34.

Williams, J. M., C. A. Duckworth, M. D. Burkitt, A. J. M. Watson, B. J. Campbell, and D. M. Pritchard. 2015. “Epithelial Cell Shedding and Barrier Function: A Matter of Life and Death at the Small Intestinal Villus Tip.” *Veterinary Pathology* 52 (3): 445–55.

Wirbel J, Pyl PT, Zych K , Kashani A, et al. 2019. “Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer”. *Nat med*. *Accepted*.

Wroblewski, Lydia E., and Richard M. Peek Jr. 2013. “Helicobacter Pylori in Gastric Carcinogenesis: Mechanisms.” *Gastroenterology Clinics of North America* 42 (2): 285–98.

Wu, Na, Xi Yang, Ruifen Zhang, Jun Li, Xue Xiao, Yongfei Hu, Yanfei Chen, et al. 2013. “Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients.” *Microbial*

Ecology 66 (2): 462–70.
Wu, Shaoguang, Ki-Jong Rhee, Emilia Albesiano, Shervin Rabizadeh, Xinqun Wu, Hung-Rong Yen, David L. Huso, et al. 2009. “A Human Colonic Commensal Promotes Colon Tumorigenesis via Activation of T Helper Type 17 T Cell Responses.” *Nature Medicine* 15 (9): 1016–22.
Xiao, E., Marcelo Mattos, Gustavo Henrique Apolinário Vieira, Shanshan Chen, Jôice Dias Corrêa, Yingying Wu, Mayra Laino Albiero, Kyle Bittinger, and Dana T. Graves. 2017. “Diabetes Enhances IL-17 Expression and Alters the Oral Microbiome to Increase Its Pathogenicity.” *Cell Host & Microbe* 22 (1): 120–28.e4.
Xie, Yuan-Hong, Qin-Yan Gao, Guo-Xiang Cai, Xiao-Ming Sun, Tian-Hui Zou, Hui-Min Chen, Si-Yi Yu, et al. 2017. “Fecal Clostridium Symbiosum for Noninvasive Detection of Early and Advanced Colorectal Cancer: Test and Validation Studies.” *EBioMedicine* 25 (November): 32–40.
Xu, Rong, Quanqiu Wang, and Li Li. 2015. “A Genome-Wide Systems Analysis Reveals Strong Link between Colorectal Cancer and Trimethylamine N-Oxide (TMAO), a Gut Microbial Metabolite of Dietary Meat and Fat.” *BMC Genomics* 16 Suppl 7 (June): S4.
Yu, Jun, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao Yi Liang, Youwen Qin, Longqing Tang, et al. 2017. “Metagenomic Analysis of Faecal Microbiome as a Tool towards Targeted Non-Invasive Biomarkers for Colorectal Cancer.” *Gut* 66 (1): 70–78.
Zeller, Georg, Julien Tap, Anita Y. Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I. Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, Rajna Hercog, Moritz Koch, Alain Luciani, Daniel R. Mende, Martin A. Schneider, Petra Schrotz-King, et al. 2014. “Potential of Fecal Microbiota for Early-stage Detection of Colorectal Cancer.” *Molecular Systems Biology* 10 (11): 766.
Zhang, Chenhong, and Liping Zhao. 2016. “Strain-Level Dissection of the Contribution of the Gut Microbiome to Human Metabolic Disease.” *Genome Medicine* 8 (1): 41.
Zhang, X., M. Rimpiläinen, E. Simelyte, and P. Toivanen. 2001. “Enzyme Degradation and Proinflammatory Activity in Arthritogenic and Nonarthritogenic Eubacterium Aerofaciens Cell Walls.” *Infection and Immunity* 69 (12): 7277–84.
Zhang, Xuan, Dongya Zhang, Huijue Jia, Qiang Feng, Donghui Wang, Di Liang, Xiangni Wu, et al. 2015. “The Oral and Gut Microbiomes Are Perturbed in Rheumatoid Arthritis and Partly Normalized after Treatment.” *Nature Medicine* 21 (8): 895–905.
Zhang, Zhigang, Jiawei Geng, Xiaodan Tang, Hong Fan, Jinchao Xu, Xiujun Wen, Zhanshan Sam Ma, and Peng Shi. 2014. “Spatial Heterogeneity and Co-Occurrence Patterns of Human Mucosal-Associated Intestinal Microbiota.” *The ISME Journal* 8 (4): 881–93.
Zitvogel, Laurence, Lorenzo Galluzzi, Mark J. Smyth, and Guido Kroemer. 2013. “Mechanism of Action of Conventional and Targeted Anticancer Therapies: Reinstating Immunosurveillance.” *Immunity* 39 (1): 74–88.
Zunino, P., C. Piccini, and C. Legnani-Fajardo. 1994. “Flagellate and Non-Flagellate Proteus Mirabilis in the Development of Experimental Urinary Tract Infection.” *Microbial Pathogenesis* 16 (5): 379–85.

Appendix 1

Coverage analysis of the V4-V5 16S rRNA primers. Percentage of entries for each bacterial phyla capable of being amplified by the primer pair used in this study using two distinct databases. The releases of the SILVA (115) and the RDP (11.2) databases used for this analysis contained, respectively, 621,948 and 2,518,232 16S rRNA sequences.

	SILVA Database	**RDP database**
Taxonomy	**Coverage (%)**	**Coverage (%)**
Acidobacteria	91.9	32.2
Actinobacteria	87.4	64.0
Firmicutes	86.5	62.2
Proteobacteria	86.1	48.6
Eubacteria	84.4	52.1
Bacteroidetes	84.1	46.8
Spirochaetes	76.1	59.3
Lentisphaerae	46.3	62.2
Verrucomicrobia	21.6	10.9

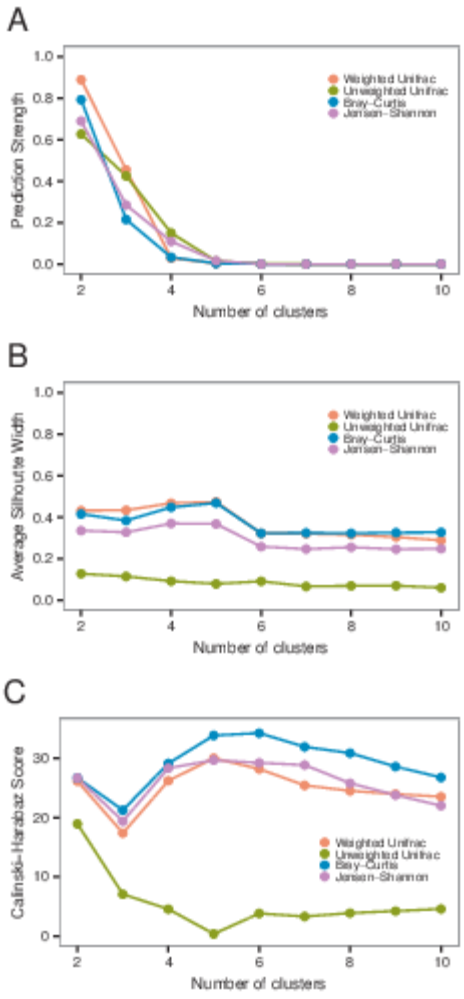
Appendix 2

ANOSIM and ADONIS P-values for beta diversity metrics.

Variable	**Bray-Curtis**	**Weighted UniFrac**	**Unweighted UniFrac**
Cancer Status	**0.001**	**0.001**	**0.001**
Age	0.174	0.084	0.416
Gender	0.39	0.467	0.765
Alcohol Use	0.837	0.616	0.821
Tobacco Use	0.443	0.317	0.511
BMI	0.328	0.346	0.215
Library Construction	0.079	0.162	0.053

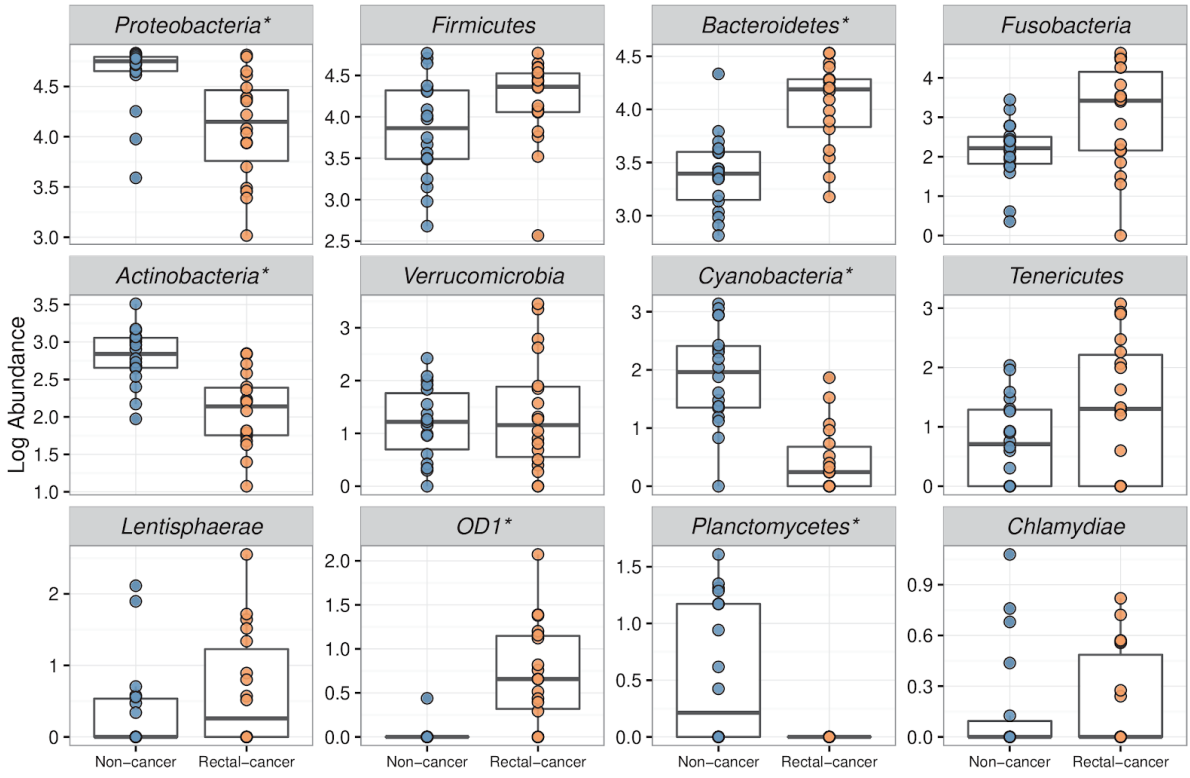
Appendix 3

Cluster structure recovered from relative abundances of genera level counts using four distance metrics. Cluster quality was tested using: **(A)** prediction strength, **(B)** silhouette index and **(C)** Caliński-Harabasz statistic.



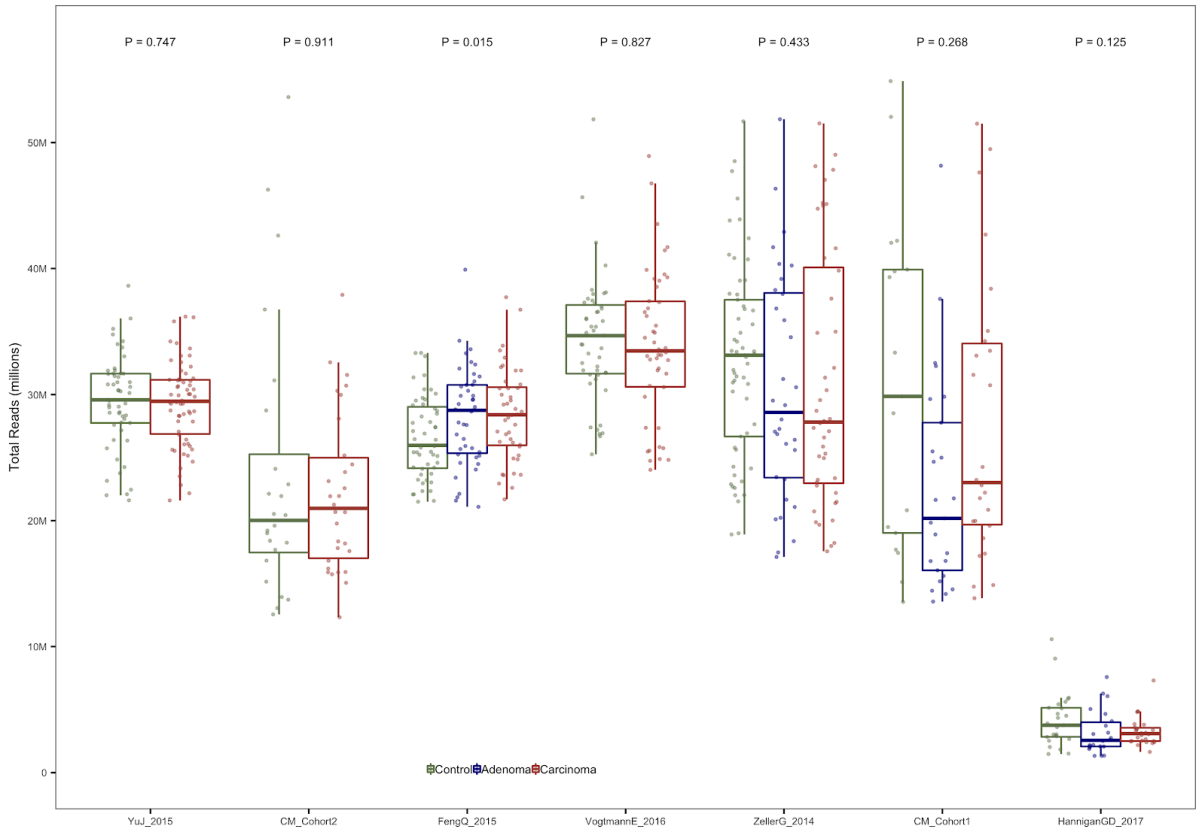
Appendix 4

Boxplots showing log abundances for the 12 most abundant phyla in rectal cancer and non cancer samples. Phyla with significant differences are labeled with (*). Significant p-values were found for *Cyanobacteria*, *Actinobacteria*, *Bacteroidetes*, *OD1* (*p-values* < 0.001), *Proteobacteria* (*p-value* = 0.002) and *Planctomycetes* (*p-value* = 0.002).



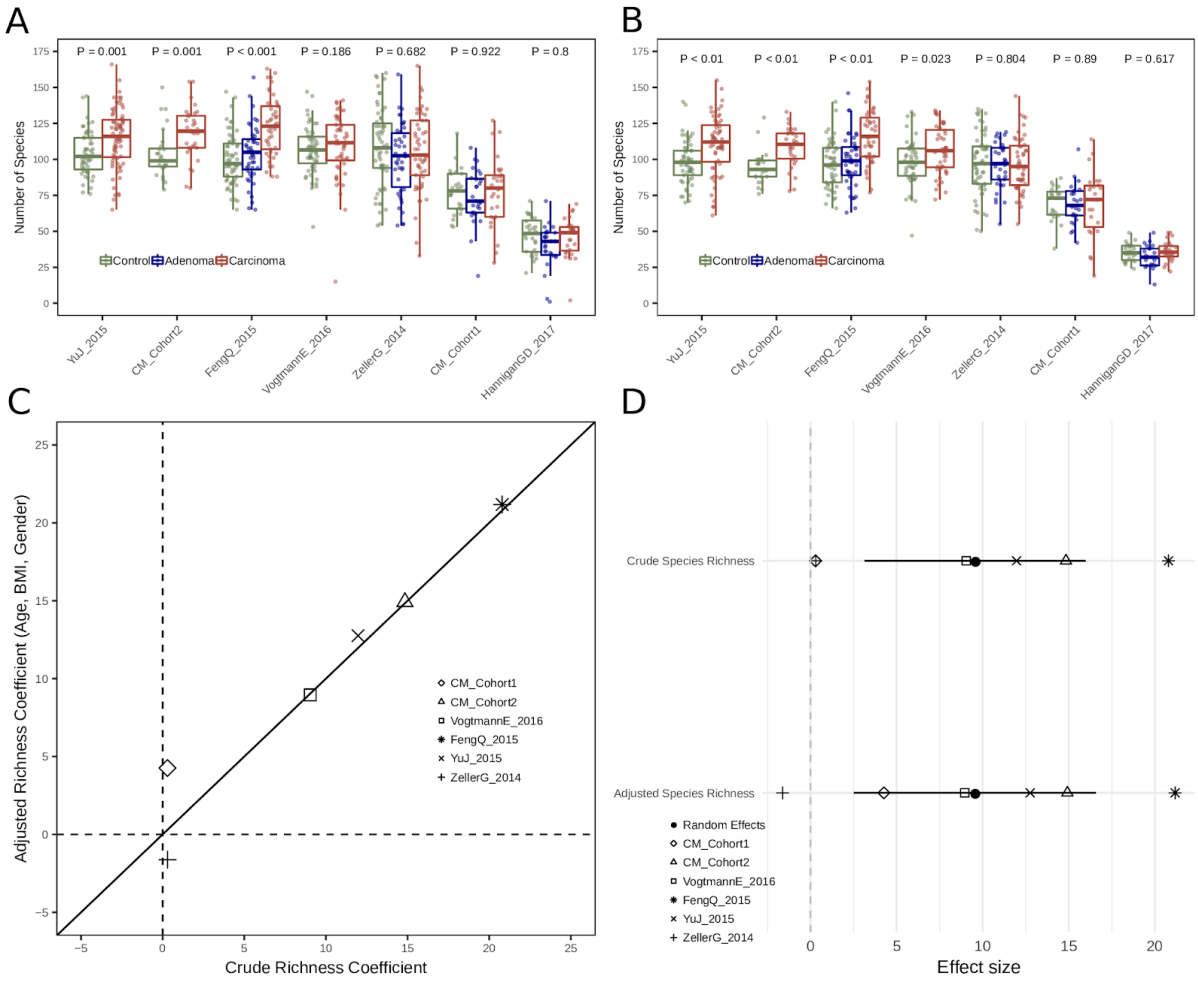
Appendix 5

Sequencing depths in CRC datasets. Boxplots showing the total number of reads per group and per dataset. P-values were calculated by Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset.



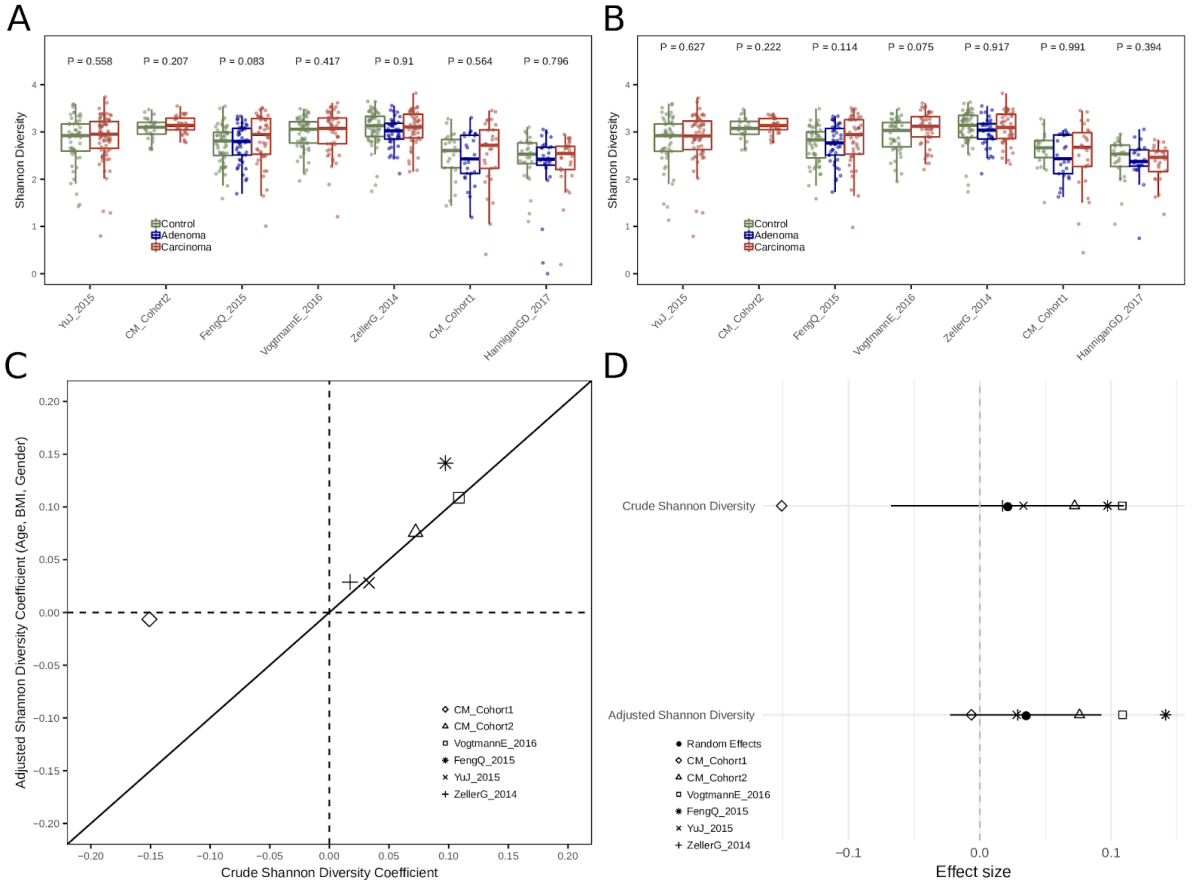
Appendix 6

Species richness in CRC datasets. (A) Boxplots showing the total number of bacterial species per group and per dataset calculated using raw metaphlan2 bowtie2 indexes. P-values were calculated by Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **(B)** Boxplots showing the total number of bacterial species per group and per dataset using rarefied metaphlan2 bowtie2 indexes to the 10th percentile. **(C)** Multivariate analysis of species richness using crude and age, gender and BMI adjusted limma coefficients. **(D)** Meta-analysis of crude and adjusted multivariate richness coefficients using a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate.



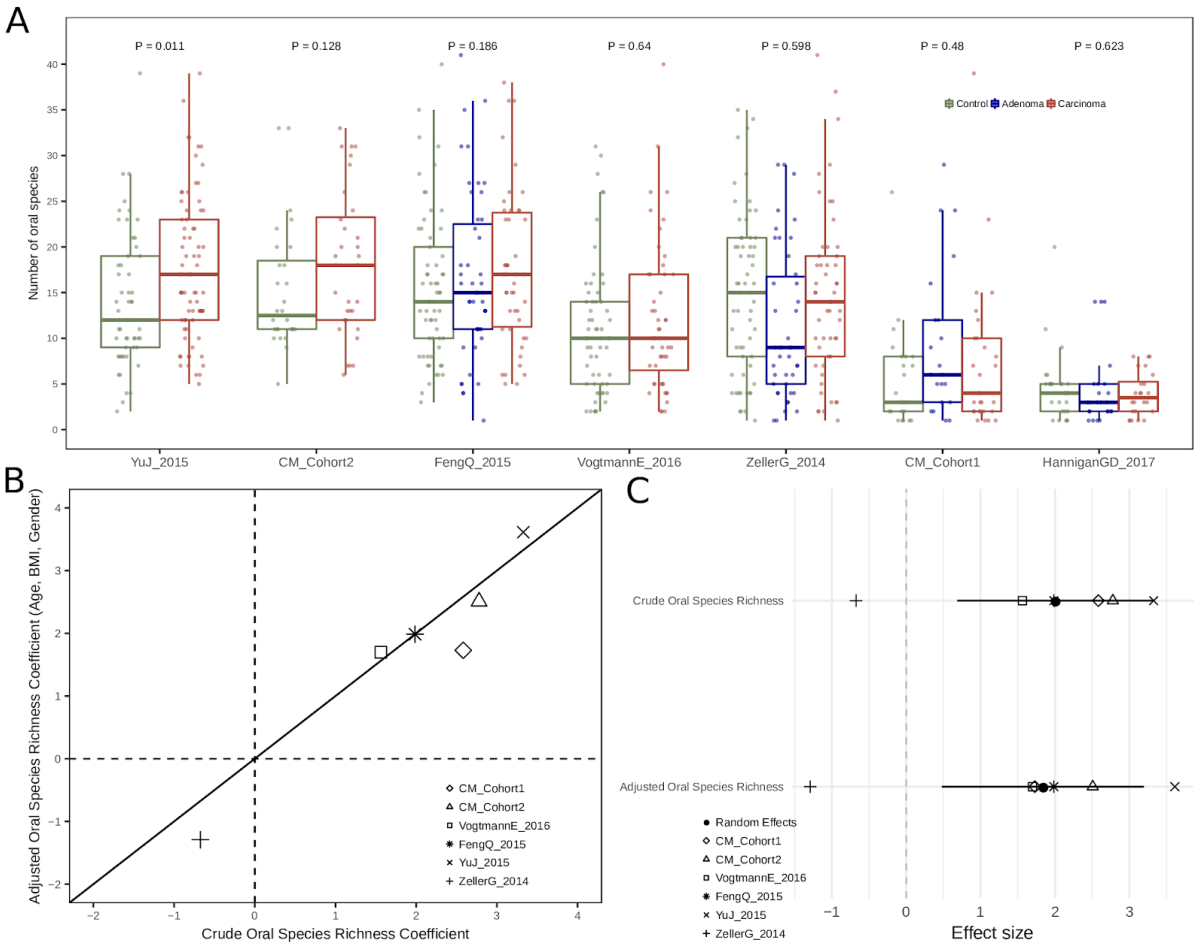
Appendix 7

Shannon diversity in CRC datasets. (A) Boxplots showing the Shannon diversity of bacterial species per group and per dataset calculated using raw metaphlan2 bowtie2 indexes. P-values were calculated by Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **(B)** Boxplots showing the Shannon diversity of bacterial species per group and per dataset using rarefied metaphlan2 bowtie2 indexes to the 10th percentile. **(C)** Multivariate analysis of species richness using crude and age, gender and BMI adjusted limma coefficients. **(D)** Meta-analysis of crude and adjusted multivariate Shannon diversity coefficients using a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate.



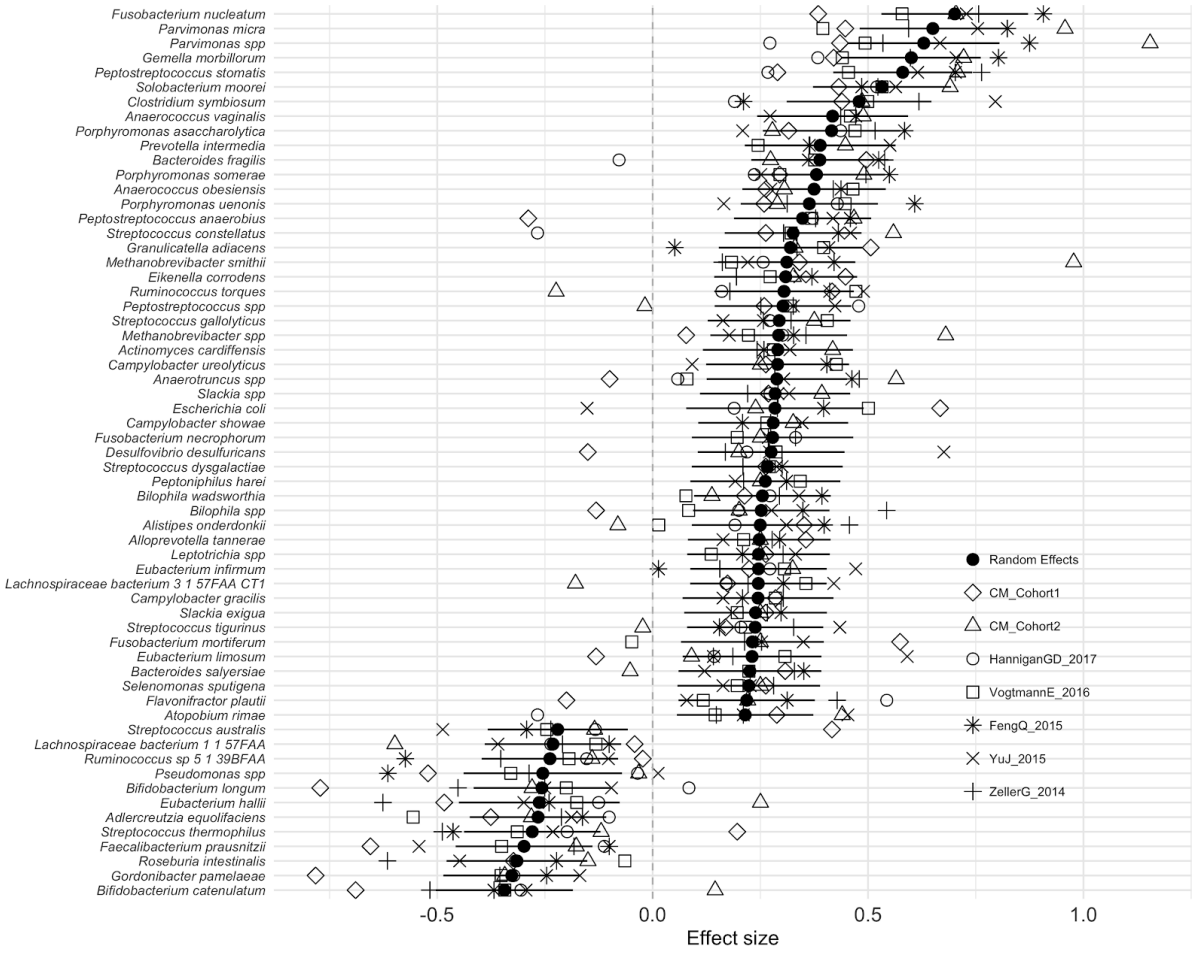
Appendix 8

Analysis of putative oral species richness in CRC datasets. (A) Boxplots showing the total number of bacterial species per group and per dataset calculated using raw metaphlan2 bowtie2 indexes. P-values were calculated by Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **(B)** Multivariate analysis of putative oral species richness using crude and age, gender and BMI adjusted limma coefficients. **(D)** Meta-analysis of crude and adjusted multivariate putative oral species richness coefficients using a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate.



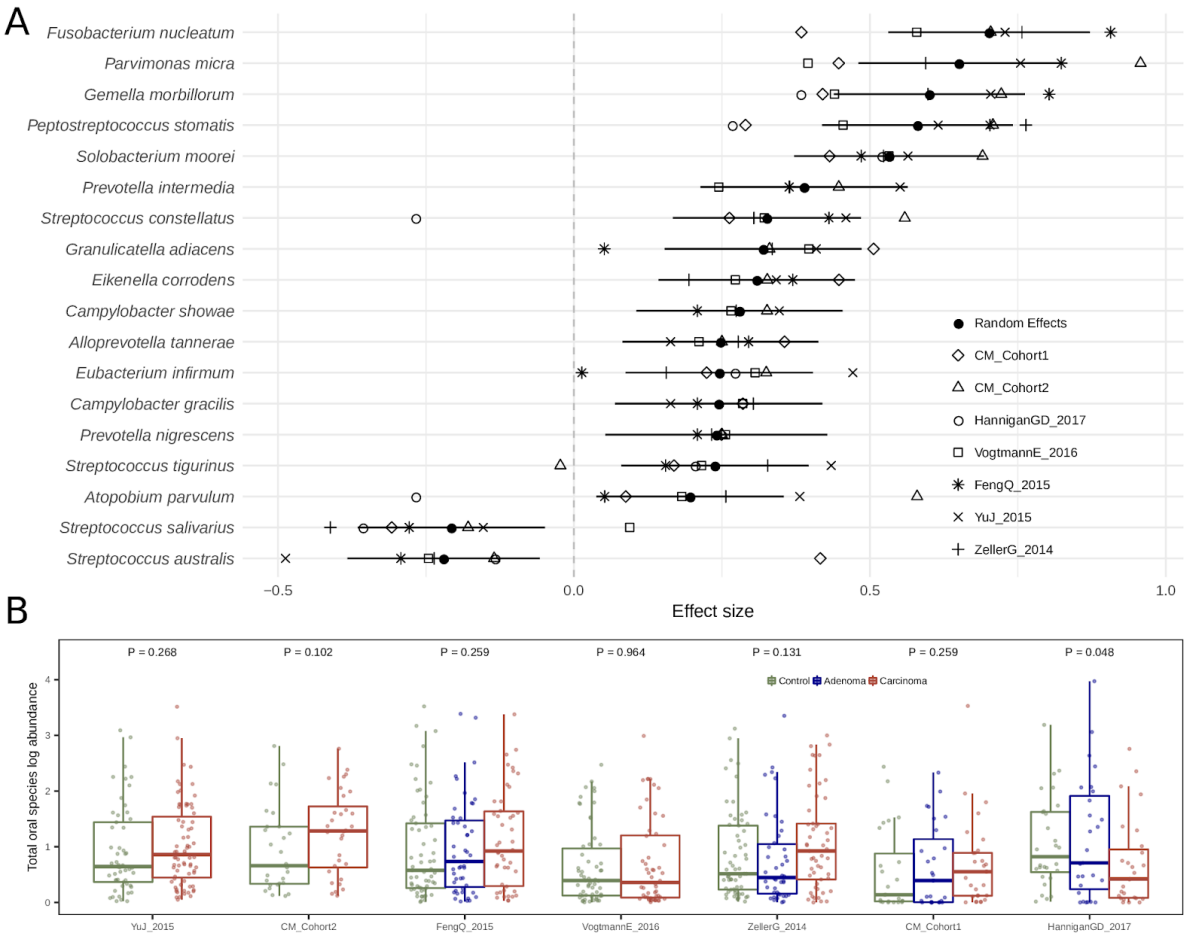
Appendix 9

Meta-analysis of CRC datasets using species-level MetaPhlAn2 profiles. Effect sizes of significant species found using a meta-analysis of standardized mean differences and a random effects model on MetaPhlAn2 species-level abundances. Bold lines represent the 95% confidence interval for the random effects model estimate.



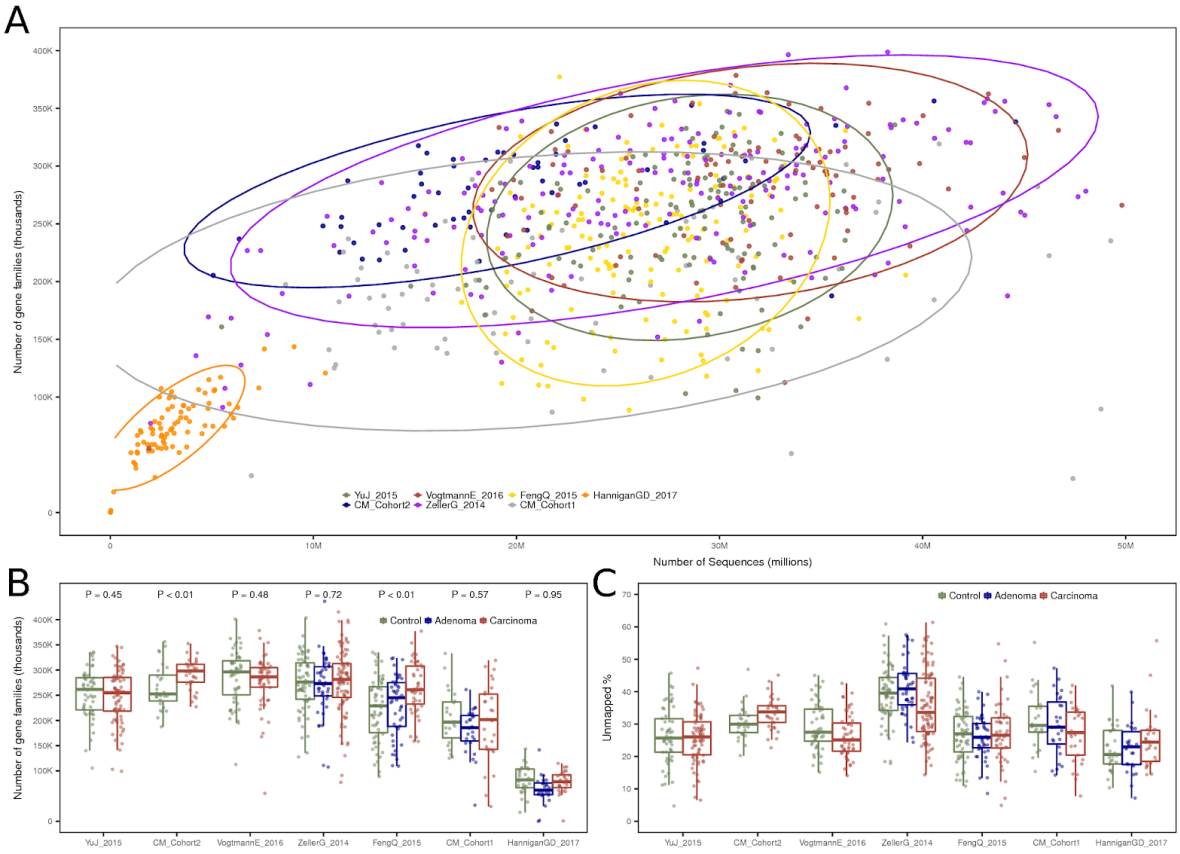
Appendix 10

Analysis of putative oral species abundances in CRC datasets. (A) Effect sizes of significant putative oral species identified using a meta-analysis of standardized mean differences and a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate. **(B)** Boxplots showing the total log abundance of putative oral species in each dataset and each group. P-values were obtained by Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset.



Appendix 11

Functional richness of gene families in CRC datasets. (A) Scatter plot showing the total number of reads and the total number of gene families identified using HUMANn2. Ellipses represent the 95% confidence level assuming a multivariate t-distribution. **(B)** Boxplots showing the total number of gene families per group and per dataset. P-values were obtained by Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **(C)** Boxplots showing the percentages of unmapped reads across datasets and groups for UniProt90 gene families.



Appendix 12

Results from StrainPhlAn associations with metadata.

		ANOSIM on distances from phylogenetic trees			Clustering of distances using PAM (only species with prediction strength > 0.8)			
Species	Sample Size	Disease *p-value*	Country *p-value*	Dataset *p-value*	Maximum Prediction Strength	Number of Clusters	Disease Fisher *p-value*	Country Fisher *p-value*
Faecalibacterium spp.	462	0.972	0.001	0.001	0.66	NA	NA	NA
*Eubacterium rectale**	375	0.951	0.001	0.001	0.96	2	0.736	0
Bifidobacterium longum	304	0.673	0.001	0.001	0.83	2	0.064	0.961
Ruminococcus torques	299	0.002	0.471	0.036	0.77	NA	NA	NA
Bifidobacterium adolescentis	241	0.604	0.153	0.04	0.62	NA	NA	NA
*Dorea formicigenerans**	225	0.997	0.001	0.001	0.92	3	0.439	0
Eubacterium hallii	224	0.223	0.019	0.015	0.88	2	0.673	0.027
Bacteroides caccae	215	0.247	0.001	0.001	0.81	3	0.234	0.02
Escherichia coli	215	0.057	0.029	0.025	0.99	2	0.139	0.101
Akkermansia spp.	214	0.638	0.001	0.005	1.00	2	0.454	0.002
Barnesiella spp.	193	0.597	0.182	0.904	0.74	NA	NA	NA
Parabacteroides merdae	192	0.623	0.332	0.373	0.96	2	0.527	0.152
Roseburia intestinalis	169	0.262	0.001	0.001	0.65	NA	NA	NA
Ruminococcus lactaris	164	0.895	0.001	0.001	0.66	NA	NA	NA
Dorea longicatena	152	0.401	0.439	0.437	0.73	NA	NA	NA
Ruminococcus obeum	149	0.331	0.638	0.622	0.66	NA	NA	NA
Eubacterium siraeum	148	0.538	0.002	0.003	0.98	2	0.71	0.002
Ruminococcus gnavus	135	0.036	0.002	0.003	0.86	3	1	0.267
Bacteroides thetaiotaomicron	132	0.993	0.909	0.887	0.80	NA	NA	NA
Streptococcus salivarius	123	0.08	0.008	0.004	0.69	NA	NA	NA
Methanobrevibacter smithii	121	0.073	0.036	0.028	1.00	2	1	0.153
Alistipes putredinis	120	0.26	0.001	0.01	0.94	2	1	0.045
Bacteroides uniformis	116	0.381	0.373	0.127	0.67	NA	NA	NA

Ruminococcus bromii	113	0.784	0.001	0.001	0.59	NA	NA	NA
Phascolarctobacterium spp.	112	0.403	0.001	0.001	1.00	2	0.678	0
Streptococcus thermophilus	105	0.449	0.24	0.104	0.74	NA	NA	NA
Coprococcus comes	100	0.86	0.001	0.001	0.95	2	1	0.049
Bacteroides dorei	96	0.952	0.014	0.329	0.69	NA	NA	NA
Bifidobacterium bifidum	87	0.85	0.016	0.107	0.61	NA	NA	NA
Coprococcus sp ART55 1	81	0.49	0.001	0.002	0.66	NA	NA	NA
Odoribacter splanchnicus	81	0.055	0.003	0.002	0.75	NA	NA	NA
Bacteroides stercoris	71	0.349	0.39	0.622	0.68	NA	NA	NA
Bacteroides eggerthii	69	0.461	0.08	0.295	0.95	3	0.568	0.602
Streptococcus parasanguinis	69	0.791	0.236	0.313	0.79	NA	NA	NA
Roseburia hominis	68	0.942	0.003	0.002	0.72	NA	NA	NA
Bacteroides massiliensis	66	0.671	0.014	0.02	1.00	2	0.77	0.07
Roseburia inulinivorans	66	0.858	0.028	0.003	0.43	NA	NA	NA
Bacteroides cellulosilyticus	64	0.309	0.419	0.176	0.62	NA	NA	NA
Bacteroides fragilis	58	0.214	0.431	0.399	0.45	NA	NA	NA
Bacteroides ovatus	55	0.436	0.686	0.613	0.94	2	0.574	0.436
Butyrivibrio crossotus	46	0.82	0.016	0.04	0.99	2	1	0.022
Dialister invisus	46	0.81	0.149	0.073	0.74	NA	NA	NA
Sutterella wadsworthensis	45	0.995	0.011	0.089	0.68	NA	NA	NA
Bacteroides salyersiae	43	0.346	0.019	0.042	0.69	NA	NA	NA
Bacteroides vulgatus	36	0.732	0.023	0.051	0.63	NA	NA	NA
Clostridium bartlettii	36	0.425	0.007	0.003	0.44	NA	NA	NA
Clostridium sp L2 50	34	0.284	0.001	0.003	0.55	NA	NA	NA
Klebsiella pneumoniae	33	0.862	0.924	0.943	0.29	NA	NA	NA
Parabacteroides distasonis	32	0.03	0.668	0.607	0.00	NA	NA	NA
Clostridium leptum	29	0.79	0.705	0.701	0.41	NA	NA	NA
Clostridium symbiosum	29	0.161	0.013	0.019	0.53	NA	NA	NA

Appendix 13

Prediction performances using MetaPhlAn2 species abundances for all datasets at increasing numbers of external datasets considered in the training model. The dark yellow line interpolates the median AUC at each number of training datasets considered.