

Computational Tools and Resources for Pan-Cancer Analyses of Host-Microbe
Interactions

by

Anders Benton Dohlman

Department of Biomedical Engineering
Duke University

Date: _____

Approved:

Lingchong You, Supervisor

Lawrence David

Jessilyn Dunn

Michael Lynch

Sayan Mukherjee

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy in the Department of
Biomedical Engineering in the Graduate School
of Duke University

2022

ABSTRACT

Computational Tools and Resources for Pan-Cancer Analyses of Host-Microbe Interactions

by

Anders Benton Dohlman

Department of Biomedical Engineering
Duke University

Date: _____

Approved:

Lingchong You, Supervisor

Lawrence David

Jessilyn Dunn

Michael Lynch

Sayan Mukherjee

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry in the Graduate School of Duke University

2022

Copyright by
Anders Benton Dohlman
2022

Abstract

The human microbiome is a dynamic, integrated ecosystem that interacts with the host to influence cancer development and progression, as well as affect response to anti-cancer therapies, suggesting opportunities for diagnostic and therapeutic approaches. Many microbe-microbe and host-microbe interactions relevant to cancer are expected to take place at the tumor site. However, obtaining and sequencing biological samples for the interrogation of these interactions is costly, while the exponential growth of sequencing data for such samples poses analytical and interpretive challenges. Thus, there is a growing need for comprehensive resources, reference databases, and analytical tools for understanding host-microbe interactions relevant to human cancers and other diseases. Herein, I demonstrate that the creation of such resources does not necessitate massive investments into new research programs, and can instead be accomplished by utilizing preexisting, public information. In two trans-kingdom, pan-cancer analyses of sequencing data from The Cancer Genome Atlas (TCGA), it is shown that both bacteria and fungi are involved human tumors samples, and that these signatures are predictive of patient outcomes. In doing so, novel methods for mitigating contamination and false-positive signals in such datasets are described. Lastly, a widely applicable analytical tool and reference database for microbe set enrichment analysis is proposed, which can be used to interpret large microbiome datasets.

Dedication

To my farfar Claes Henrik Dohlman. His lifelong devotion to science has been a profound inspiration to me.

Contents

Abstract	iv
List of Tables	xii
List of Figures	xiii
List of Abbreviations	xv
Acknowledgements	xvii
1 Introduction.....	1
1.1 Human Interactions with Microorganisms	1
1.2 The Human Microbiome and Cancer	3
1.3 Understanding the Microbiome via Publicly Available Data	7
1.4 Organization of the Dissertation	10
2 Mapping the Microbial Interactome: Statistical and Experimental Approaches for Microbiome Network Inference.....	12
2.1 Introduction	12
2.2 Ecology Meets Network Theory.....	16
2.3 Cross-Sectional Methods for Detecting Microbial Interactions	20
2.4 Longitudinal Methods for Detecting Microbial Interactions	27
2.5 The Expanded Universe of Microbial Interactions.....	34
3 The Cancer Microbiome Atlas: A Pan-Cancer Comparative Analysis to Distinguish Organ-Associated Microbiota from Contaminants.....	41
3.1 Introduction	41
3.2 Results.....	47

3.2.1	WGS and WXS harbor colorectal bacterial reads distinct from blood and brain	47
3.2.2	Species equiprevalent in tissue and blood are predominantly contaminants..	50
3.2.3	Equiprevalent species are associated with particular sequencing centers....	53
3.2.4	A generalizable model for isolating tissue-resident microbiota in TCGA tumor samples	55
3.2.5	<i>Proteobacteria</i> and <i>Actinobacteria</i> contribute the largest fraction of contaminant reads	56
3.2.6	Detecting tissue-resident and contaminant species with gene-level resolution.....	57
3.2.7	Distinguishing tissue-resident <i>Escherichia</i> reads from contamination.....	59
3.2.8	Tissue-enriched sequencing reads can be identified with nucleotide precision.....	61
3.2.9	Decontamination removes sequencing center artifacts	62
3.2.10	Original TCGA tissue and blood samples validate tissue-resident microbial compositions and equivalent species as contaminants.....	65
3.2.11	Colorectal tissue microbiomes cluster into <i>Fusobacterium</i> and <i>Bacteroides</i> coabundance groups	68
3.2.12	Bacterial coabundance groups are predictive of the host tissue molecular environment	69
3.2.13	Matched tumor-normal analysis reveals known and novel species associated with colorectal neoplasms.....	71
3.2.14	Survival analysis reveals candidate microbial biomarkers predictive of clinical outcomes	72
3.2.15	Microbial presence in CRC tissue is predictive of host immunogenic response, inflammatory cancer pathways, and cell-cell adhesion	73

3.2.16	Microbial presence in CRC blood samples indicate mucosal barrier injury.	76
3.2.17	Contamination-adjusted tissue microbiome profiles for all gastrointestinal cancers in TCGA.....	79
3.3	Discussion.....	80
3.4	Methods	83
3.5	Acquisition and metagenomic profiling of TCGA sequencing data.....	83
3.5.1	Decomposition of observed TCGA microbial profiles into tissue-resident and contaminant fractions.....	84
3.5.2	Gene-level sequencing analysis of representative species	87
3.5.3	Nucleotide-level analysis of bacterial sequence variants	88
3.5.4	Acquisition and analysis of original TCGA tissue and plasma samples.....	88
3.5.5	Estimation of bacterial coabundance groups and associated molecular signatures	90
3.5.6	Identification of tumor- and normal tissue-associated microbiota	91
3.5.7	Survival analysis.....	92
3.5.8	Pathway analysis of species associated with tumors or adjacent normal tissue.....	92
3.5.9	Quantification and statistical analysis.....	93
4	A pan-cancer mycobiome analysis reveals fungal involvement in gastrointestinal and lung tumors that is predictive of survival	94
4.1	Introduction	94
4.2	Results.....	100
4.2.1	Fungal DNA is abundant in gastrointestinal tumor samples from TCGA .	100

4.2.2	Identification and removal of contaminant fungi and false-positive signals ...	100
4.2.3	TCGA tissue samples are composed of disease-specific fungi	107
4.2.4	Emergence of <i>Candida</i> and <i>Saccharomyces</i> co-abundance groups is associated with gastrointestinal cancers	111
4.2.5	Trans-kingdom analysis reveals co-abundance groups associated with <i>Candida</i> and <i>Saccharomyces</i> in GI cancers	111
4.2.6	<i>Candida</i> and <i>Saccharomyces</i> are predictive of gene expression patterns in GI cancers	112
4.2.7	A <i>Candida</i> -to- <i>Saccharomyces</i> ratio is associated with late-stage, metastatic colon cancer.....	117
4.2.8	Analysis of blood samples suggests fungal translocation from GI tumors to the bloodstream	Error! Bookmark not defined.
4.2.9	Targeted analysis of <i>Candida</i> and <i>Saccharomyces</i> spp.	123
4.2.10	Cancer-associated mycobiota and clinical outcomes highlight predictive value of <i>Candida</i>	127
4.3	Discussion.....	130
4.4	Methods	133
4.4.1	Detection and quantification of mycobiota in TCGA sequencing data	137
4.4.2	Identification and removal of contaminant fungi in TCGA sequencing data ..	137
4.4.3	Quality control by vertical and horizontal analyses of fungal genome coverage.....	141
4.4.4	Validation with TaxaTarget	Error! Bookmark not defined.
4.4.5	Targeted analysis and quantification of <i>Candida</i> and <i>Saccharomyces</i> species of interest.....	143

4.4.6	Estimation of intra- and inter-kingdom co-abundance groups and associated gene expression signatures	143
4.4.7	Quantification of live <i>Candida</i> in primary colorectal tumor samples	143
4.4.8	Identification of <i>Candida</i> - and <i>Saccharomyces</i> -type TCGA tumor samples and associated signatures	146
4.4.9	Differential abundance analysis between tumor and adjacent normal tissue	146
4.4.10	Survival analysis.....	148
4.4.11	Random forest classification of cancer types using fungal compositions of tumor and blood samples.....	149
5	Taxonomic Set Enrichment Analysis: A Curated Database and Analytical Toolkit for Interpreting Metagenomic Data.....	151
5.1	Introduction	151
5.2	Methods	155
5.2.1	Continuous and discrete TaxSEA	155
5.2.2	An ontology for microbial traits.....	157
5.2.3	Creation of a microbial traits database.....	161
5.2.4	An interactive website for interpreting microbiome data	162
5.3	Results.....	163
5.3.1	TaxSEA accurately identifies human body sites.....	163
5.3.2	Microbiome features associated with colorectal tumors	165
5.4	Discussion.....	169
6	Conclusion.....	172
	References	177

Biography 208

List of Tables

Table 2.1: Summary of statistical tools for inference of weighted, undirected microbial interaction networks from cross-sectional metagenomic data	23
Table 2.2: Summary of statistical tools for inference of directed microbial interaction networks from longitudinal metagenomic data	28
Table 5.1: Hierarchy of microbial traits and associated sources.....	158

List of Figures

Figure 2.1: The Cartesian plane of pairwise ecological interactions.....	18
Figure 2.2: Network abstractions of the microbial interactome.....	35
Figure 3.1: WGS and WXS harbor colorectal bacterial reads distinct from blood and brain	46
Figure 3.2: Most equiprevalent taxa are common contaminants and associated with particular sequencing centers.....	52
Figure 3.3: Detecting tissue-resident and contaminant species with gene-level resolution	58
Figure 3.4: Decontamination removes sequencing center artifacts and original TCGA tissue and blood samples validate tissue-resident microbial compositions and equivalent species as contaminants	63
Figure 3.5: Colorectal tissue microbiomes cluster into Fusobacterium and Bacteroides coabundance groups predictive of host tissue molecular environment	67
Figure 3.6: Microbial presence in CRC tissue is predictive of host gene expression pathways and mucosal barrier injury	75
Figure 3.7: Contamination-adjusted tissue microbiome profiles for all gastrointestinal cancers in TCGA	78
Figure 4.1: Fungal DNA is present in multiple cancer types not explained by contamination.....	99
Figure 4.2: Primary tumor samples harbor disease-specific mycobiomes.....	106
Figure 4.3: Trans-kingdom analysis reveals Candida- and Saccharomyces-associated GI cancer coabundance groups	110
Figure 4.4: Candida is associated with late-stage and metastatic GI cancers.....	Error! Bookmark not defined.
Figure 4.5: Fungal DNA is detected in the blood of cancer patients and may enter the circulation through mucosal barriers	Error! Bookmark not defined.

Figure 4.6: Candida species are present in GI cancers and high abundance is associated with early-stage stomach cancer	126
Figure 4.7: Cancer-associated fungal mycobiota and clinical outcomes highlight predictive value of Candida	129
Figure 5.1: Cartoon depicting the organization of the TaxSEA database	157
Figure 5.2: PRISMA diagram depicting the selection process for including microbiome resources in the TaxSEA database	160
Figure 5.3: Screenshot of the TaxSEA website	163
Figure 5.4: TaxSEA correctly identifies host body sites from an independent reference	164
Figure 5.5: TaxSEA links colorectal cancer to periodontitis.....	167
Figure 5.6: TaxSEA identifies metabolic functions and potential treatments for CRC microbiomes.....	168

List of Abbreviations

- BC = Brain Cancer
- BRCA = Breast Invasive Carcinoma
- CEA = Continuous Enrichment Analysis
- COAD = Colorectal Adenocarcinoma
- CRC = Colorectal Cancer
- DEA = Discrete Enrichment Analysis
- ESCA = Esophageal Carcinoma
- GBM = Glioblastoma Multiforme
- GI = Gastrointestinal
- GSEA = Gene Set Enrichment Analysis
- HMP = Human Microbiome Project
- HNSC = Head-Neck Squamous Cell Carcinoma
- LGG = Brain Lower Grade Glioma
- LUSC = Lung Squamous Cell Carcinoma
- MSEA = Microbe Set Enrichment Analysis
- READ = Rectal Adenocarcinoma
- STAD = Stomach Adenocarcinoma
- TaxSEA = Taxonomic Set Enrichment Analysis
- TCGA = The Cancer Genome Atlas

- TCFA = The Cancer Fungi Atlas
- WGS = Whole Genome Sequencing
- WXS = Whole Exome Sequencing

Acknowledgements

First, I would like to acknowledge my family, who have been supportive to me in so many ways throughout my PhD. I am so happy to have lived close to my parents during these past five years in Durham. Thank you for the delicious meals and stimulating conversations.

To my dear friends, both near and far: Sam, Darien, Isaac, Andrew, Eriq, Willa, Lex, Christian, Aeran, Sejiro, Robert, Jocelyn, Charley, and Matt, thank you for all the good times and much-needed distractions.

To my committee members, thank you for your kind support and constructive feedback over the years.

Finally, I would like to thank my graduate advisor and mentor, Xiling Shen who provided valuable guidance throughout my PhD. He took a chance recruiting me, and I will forever be grateful for the opportunity.

1 Introduction

1.1 *Human Interactions with Microorganisms*

Microorganisms are inextricably bound to humankind, having coevolved with both our internal biology [1, 2] and our external environment [3]. Microbiota have been interwoven with the fabric of civilization, even prior to their discovery. *Saccharomyces cerevisiae* was unconsciously used by Egyptians to develop processes of brewing and baking as many as 7000 years ago, with the earliest evidence of intentional fermentation being around 13,000 years ago. Meanwhile, pathogenic microorganisms and epidemics of infectious disease have profoundly influenced the course of humanity to this day, as well as the course of human evolution [1, 2]. In addition to having their own immune systems, microorganisms have played a critical role in the evolution of multicellular organisms' own immune systems through a so-called "evolutionary arms-race" [2].

However, microbiota as we know them now were not discovered until the 17th century, following improvements to the microscope made by Robert Hooke [4] and Antonie van Leeuwenhoek. Athanasius Kircher may have been the first to conclude that microorganisms were the origin of disease, describing "little worms" in the blood of plague victims in Rome [5]. However, the notion that microbiota were the causative agents of many diseases became well-accepted by the turn of the 18th century, leading to many important therapeutic and theoretical developments in the field of medicine. In

the modern era, microorganisms remain crucial for public health [6], agriculture [7], industrial production [8], and scientific development [9, 10].

Like the discovery of the microscope, it was methodological developments that ushered in a dramatic expansion in our understanding of the scope of interactions between humans and microbiota. The discovery of nucleic acids and the development of methods for DNA sequencing have allowed culture-independent studies of bacterial diversity, dramatically expanding the universe of known microorganisms; it is estimated that more than 99% of all microorganisms in nature are only observable by such methods [11, 12]. As such, metagenomics sequencing now allows the simultaneous profiling of entire communities of microorganisms, leading to a dramatic increase in the number characterized microbial genomes as well a newfound appreciation for the role of human microbiome in human health and disease. Between 2007 and 2016, the Human Microbiome Project (HMP) [13] set out to characterize the diversity and function of the microorganisms that colonize both interior and exterior surfaces of human body, identifying distinct ecosystems at the skin's surface as well as at oral, vaginal, and gastrointestinal sites. Today, it is estimated that the human body supports an ecosystem of 10-100 trillion microorganisms, accounting a ratio of roughly 1:1 bacterial cells per human cell [14].

Since the HMP began, the human microbiome has been implicated in a broad array of processes relevant to health. These include gastrointestinal and metabolic conditions for which the role of the microbiome is perhaps unsurprising: dysbiotic microbiomes have been implicated in the pathology of diseases such as obesity [15], diabetes [16], and IBD [17, 18]. More perplexing however are the discoveries of enigmatic links between the microbiome and neurological conditions and disorders such as depression, anxiety [19], schizophrenia [20], autism [21], and Parkinson's [22]. As identifiable connections between the human microbiome and health conditions become more numerous, there is an increasing acceptance that host-associated microorganisms are not merely uninvolved passengers that exist independently of human physiology, but rather play an active role in human health and disease. Moreover, such developments suggest that the microbiome has considerable potential as a target for therapeutic intervention and non-invasive diagnostics.

1.2 The Human Microbiome and Cancer

Clinical evidence linking the microbiome to cancer date back millennia. The Ebers Papyrus, an Egyptian medical document dated to 1550 B.C., suggests what must be described as an early use of immunotherapy: a treatment for tumors that involves introducing infection through application of a poultice, followed by an incision [23]. Over the centuries, several other reports have described the amelioration of human

cancers by microbial infections, whether introduced intentionally or otherwise [24, 25]. While reports linking bacteria and fungi to human cancer remained unsubstantiated [26] for some time, the 1911 discovery of the oncogenic Raus sarcoma virus represented the first report to identify a microorganism that was incontrovertibly capable of inducing cancer development [27]. Inability to reproduce this finding with other cancers meant the consideration of alternative hypotheses. Although the somatic mutation theory has been well established for some time, the role of the microbiome in cancer has received increasing attention over the last decade. This growing awareness stems from reports linking microorganisms to cancer development, progression, and outcomes, as well as indications that they may be used as diagnostics or influence cancer treatment [28-39].

A total of 11 species have now been designated as “oncomicrobes” by the International Association for Cancer Registries (IACR) and are expected to cause an estimated 2.2 million cases of global cancer cases per year, around 13% [40]. A large fraction of these cases are related to *Helicobacter pylori*, which is responsible for approximately 75% of attributable risk for gastric cancer [41]. However for colorectal cancer (CRC), substantial evidence has identified mechanistic roles for colibactin-expressing (pks+) *Escherichia coli* [29], enterotoxigenic *Bacteroides fragilis* [42], and *Fusobacterium nucleatum* [28, 39, 43, 44] in patient outcomes and therapeutic response. Links between microbiota and cancer have also been identified in other gastrointestinal

sites of high microbial biomass, such as oral [45], esophageal [46, 47], and the stomach [41]. However, a cancer-specific microbiome has also been identified in tumors from tissues with lower microbial load [35], suggesting the presence of tumor-involved microbiota in cancers such of the breast [48], lung [49], ovary [50], pancreas [31], and in melanoma [30, 34].

The mechanisms for cancer-microbe interactions are wide-ranging. Broadly, human microbiota and microbiota-derived metabolites affect host adaptive immunity and immune homeostasis through extensive interactions with B cells and T cells [51], which is expected to account for the numerous associations between the commensal microbiota and response to anti-cancer therapies [33, 52, 53]. In particular, specific microbiota have been shown to stimulate capable of stimulating Th17 and Th1 response to reduce tumor growth [32, 53] and can prime T cells against melanoma [33]. At a local level, microbiota affect the tumor microenvironment through both immunosuppressive and immunostimulatory mechanisms, which can shape response to immunotherapy [35]. Finally, microbiota interact extensively with the epithelial barriers at which many tumors develop [54-56]. In the lower intestine, dysbiotic and pro-inflammatory microbiota may encourage epithelial-to-mesenchymal transition microbiota [54, 55], leading to increased tight junction permeability, a significant risk factor for metastasis

[56]. For non-gastrointestinal sites, compositional changes in the commensal microbiota of the lungs and skin may contribute to carcinogenesis [38, 57].

Naturally, these observations have led to proposals for leveraging the microbiome to enhance cancer treatment. Antibiotics have been used extensively for treating gastric cancers with *H. pylori* involvement and may also help to provoke an anti-tumorigenic immune phenotype by eliminating intratumoral microbiota in lung, colon, and pancreatic cancers [28, 37, 38]. Non-pharmaceutical interventions such as diet, which has long been known to influence cancer [58], as well as prebiotics, postbiotics, and probiotics may be leveraged to influence the gut microbiota and consequently, tumor microbiomes or tumor metabolism. The introduction of beneficial, anti-tumor microbiota might also be accomplished via fecal-microbiome transplant (FMT) or through engineered, exogenous bacteria [25].

Beyond treatment, there is evidence that the microbiome might also be used as an indicator of disease. Because the microbiome can be assayed non-invasively through stool samples, swabs, or even by blood draw, there is significant interest in its prognostic and diagnostic capabilities. For example, it has been shown that colorectal cancer patients have distinct stool microbiomes [59], while bacteremia by *Streptococcus gallolyticus* can predict colorectal cancer type and location [60]. Microbiome-based cancer diagnostic strategies generally rely on sequencing technology and have been

successfully tested on cancers affecting the respiratory and gastrointestinal tracts [36, 61]. Analysis of blood and tumor samples from other cancer types suggest the presence of distinct microbiomes at other cancer types as well, which might be leveraged for diagnostic applications [35, 62].

Indeed, a substantial body of research provides clear evidence that the microbiome is a crucial area of focus if we wish to comprehensively understand human cancer development and progression. In the most recent update to his famous “Hallmarks of Cancer” series, Douglas Hanahan’s highlights polymorphic microbiomes as a major contributor to characteristics that facilitate cancer development [63]. Thus, the microbiome should be considered a powerful asset for treating and diagnosing human cancers. However, future work will be needed to better understand these links and operationalize microbiome-based diagnostics and therapies.

1.3 Understanding the Microbiome via Publicly Available Data

In December of 1982, the first DNA sequences were uploaded to GenBank, together representing fewer than one megabase of nucleotide letters. As of this writing, GenBank now holds more than 18,000 gigabases of sequence representing 250,000 unique organisms, having doubled about every 18 months. Of course, GenBank represents only a tiny fraction of the entire universe of sequencing data; the true number of sequenced nucleotides likely exceeds 10^{21} [64]. This colossal, exponential growth in

sequencing data has in large part been allowed by advancements in both sequencing technology and computing. As the availability of public sequencing data has grown, so too has the ecosystem of analytical tools and statistical methodologies used to analyze such datasets [65]. Today, meaningful scientific analyses can be performed using entirely publicly available data [66], without the need for expensive cell lines, mouse colonies, or chemical reagents.

While the first applications of new sequencing technologies were largely developed with the human genome in mind, recent years have been characterized by a substantial increase in studies on microbial genomics. Deemed the “spiritual successor” to The Human Genome Project (HGP), the HMP [13] inspired decades of substantial research into the role of microorganisms in human health and disease. Improvements in the availability of metagenomic reference databases and statistical methodology for interpreting metagenomics results have greatly enhanced the tools at our disposal for microbiome research and its applications to a broad array of fields, including medicine, biotechnology, ecology, and public health. An enormous body of literature exists for microorganisms, spanning several databases and millions of published research articles from both individual research laboratories and multi-institutional sequencing initiatives. To that end, meta-analyses of the available literature can be used to extract information

on microbiota which can be compiled for the analysis and interpretation of metagenomic data [67-70].

Datasets of sequenced biological samples harbor troves of metagenomic information which can be mined to study microbe-microbe and host-microbe interactions [62, 71, 72]. Tools like PathSeq [71], Kraken2 [73], and MetaPhlAn [73] can characterize the microbial component of sequenced human tissue with precision. However, extreme care must be taken during metagenomic analyses of these datasets, particularly for body sites with low microbial biomass. Compared to swabs and stool samples, bulk tissue samples contain relatively fewer bacterial reads and are thus likely to contain a greater proportion of contaminant bacteria [74], which can originate from many sources in the laboratory environment, such as from nucleic acid extraction kits [74-77]. Aside from contamination, sequence similarity between microorganisms can easily lead to reporting of false-positive signals. Indeed, there is suspicion that contamination may dominate the results of many published microbiome studies [74]. Therefore, in analyses of low-biomass tissue samples controlling for contamination and related batch effects is an important step that must precede downstream analyses of interactions between the host tissue and the tissue-resident microbiome. Nevertheless, secondary analyses of existing tissue sequencing data have led to many interesting studies, particularly using public sequencing data from The Cancer Genome Atlas

(TCGA) [36, 39, 62]. Such approaches are particularly effective if they thoroughly address contamination and are supplemented by convincing experimental evidence, including external validation by metagenomics methods, histology, and tissue-culture.

1.4 Organization of the Dissertation

This dissertation is composed of six chapters. Of these chapters, two are direct reproductions of previously published work, one is a reproduction of work currently under peer review, and one is a work from a manuscript in preparation for publication. For better or worse, all the work presented here represents my own, original writing. Following the introduction, Chapter 2 reviews statistical and experimental strategies for identifying and analyzing microbe-microbe and host-microbe interactions. *This chapter is a direct reproduction of a review article, published in Experimental Biology and Medicine, co-authored with Xiling Shen.* Chapter 3 of the dissertation describes a pan-cancer analysis of the bacterial microbiome of gastrointestinal tissue and blood samples, and novel methodology for identifying and removing contamination. *This chapter is a direct reproduction of a research article published in Cell Host & Microbe, co-authored with Diana Arguijo Mendoza, Shengli Ding, Michael Gao, Holly Dressman, Iliyan D. Iliev, Steven M. Lipkin, and Xiling Shen.* Chapter 4 describes The Cancer Fungi Atlas (TCFA), a similar pan-cancer analysis of the fungal microbiome of gastrointestinal, lung, and breast tumors. *This chapter is a direct reproduction of a research article currently under revision at*

Cell, co-authored with Jared Klug, Marissa Mesko, Iris H. Gao, Steven Lipkin, Xiling Shen, and Iliyan D. Iliev. Chapter 5 of this dissertation describes TaxSEA, a statistical tool and reference database for performing microbe-set enrichment analysis (MSEA). *This chapter is an unpublished manuscript, co-authored with James Zheng and Xiling Shen.* I would like to sincerely thank each of my co-authors for their valuable contributions to these works. It has been an honor working with each of them.

2 Mapping the Microbial Interactome: Statistical and Experimental Approaches for Microbiome Network Inference¹

2.1 Introduction

By the time the initial phase of the Human Microbiome Project (HMP) drew to a close in 2014, it had become widely accepted that the human gut microbiome plays a dramatically underappreciated role in human health and disease. Similar international projects by the Beijing Genomics Institute, the American gut project, and the EU-funded MetaHIT have also punctuated growing worldwide interest in the human microbiome. The insights gained from these projects, along with advances in immunology, high-throughput metagenomic sequencing, and the development of statistical and computational tools for processing these data, have made large-scale analysis of microbial communities possible in a way they have never been before, spawning considerable excitement in the microbiome among scientists and nonscientists alike. Sometimes referred to as the “last organ” [78] or the “forgotten organ” [79], the human microbiome could be considered one of the last active frontiers of human physiology [80].

¹This chapter is exactly reproduced from a review article of the same name authored by A.B. Dohman and X. Shen, published in *Experimental Biology and Medicine*.

Recent work has drawn fascinating connections between changes in human microflora and a breadth of human diseases and conditions. Microbiota have been shown to play an important role in gastrointestinal and related diseases such as obesity [81, 82], diarrhea [83], diabetes [84], Irritable Bowel Syndrome (IBS) [85], Inflammatory Bowel Disease (IBD) [86, 87], and colorectal cancer [88]. Even more surprising, researchers are beginning to identify unexpected associations between the gut microbiome and neurological disorders such as autism [89, 90], schizophrenia [91], Parkinson's disease [92], as well as depression [93]. While in many cases, the mechanisms of such associations remain murky, there are indications that therapeutic interventions such as fecal transplants and probiotics may be effective in reducing the symptoms of many of these disorders [94-96]. Yet in clinical trials, many probiotics have failed to produce positive results, for conditions including eczema [97], diarrhea [98], and gastroenteritis [99, 100]. Critically absent from the design of these clinical trials is an adequate understanding of how probiotic therapies affect the microbiome on a systems-level, which would ostensibly guide species selection, dosing regimens, and even the engineering of healthier gut microbiomes.

Put another way, our current knowledge of both commensal and pathogenic microbes remains primarily restricted to pairwise interactions. While the behaviors and mechanisms of specific organisms have become well documented, a thorough

characterization of the multispecies interaction network and its dynamics remains elusive. In addition to providing valuable insight into the biological function and significance of specific species, a more comprehensive and quantitative map of microbiome interactions will lead to a more detailed and systemic understanding of the ways that shifts in the composition of the microbiome can shape human health. Such knowledge will facilitate the identification of novel therapeutic interventions and inform the rational design of treatment regimens. Ultimately, a complete and quantitative understanding of the gut microbiome's interaction dynamics will allow more precise manipulations, with the goal of engineering healthcare solutions to microbiome-associated diseases. It is therefore important to define the behavior and function of the human microbiome using a systems-biology approach, by refocusing experimental and analytical strategies on multivariate interactions between species.

The mapping of many canonical human gene pathways was established through careful experimentation over several decades, and these have been validated through computational modeling, network inference, and other tools from systems biology. Such networks have been constructed effectively using a variety of established statistical approaches, such as Bayesian networks, neural networks, and graphical gaussian models [101, 102]. Yet due to inherent differences in the way that microbial survey data are collected and reported, many of these strategies have proven inadequate or

inapplicable in the context of microbial network inference. This is due in part to the fact that the number of reads identified by 16S or shotgun metagenomic sequencing vary independently of overall microbial abundance. As a result, microbiome data for a particular sample are typically presented as relative abundances, or “compositions” which sum to one [103, 104]. Additionally, because microbial content varies between samples and the abundance of some microorganisms are often below the limit of detection, microbiome sample data contain a large portion of zeros, and are therefore highly sparse. These characteristics of microbiome survey data mean that standard methods of analyzing multivariate data are likely to be ineffectual and statistically untenable [105-107].

Although the microbiome field has seen experimental methods, computational tools, and available data proliferate enormously over the last decade, statistical and experimental methods for microbial network inference remain under active development. As these networks are developed, a more comprehensive understanding of the gut microbial ecosystems will emerge, providing new opportunities for precisely and predictably altering the human microbiome. In this mini-review, we will summarize various statistical and experimental approaches to mapping and analyzing microbial interaction networks. In doing so, we will discuss some of the prominent challenges and

directions for improvement that must be considered as the field of systems microbiology develops.

2.2 Ecology Meets Network Theory

Ecological relationships in the microbial interactome can be generalized using network theory, a set of mathematical concepts describing relationships between discrete entities. A network essentially consists of a set of “nodes”, which are interconnected by “edges”. Applied to microbial ecology, the nodes of a microbial interaction network represent species or operational taxonomic units (OTUs), while the edges denote functional interactions between them. Although the mechanisms of these microbial interactions can be extraordinarily complex, they can still be characterized using familiar ecological terminology.

The nature of an ecological relationship between microbes is typified by the harmful or beneficial growth-rate effect that each microbe has on its interaction partner. Microbes can have a net negative or positive impact on one another by producing or consuming resources, but also by manipulating their environment, such as through modulations in pH [108]. Microbes competing for metabolites and macromolecules have a mutually negative effect on one another (competition), while interaction partners producing mutually beneficial metabolites or environmental conditions both benefit (mutualism). They can also exhibit opposite effects on one another, such as in predator-

prey relationships, in which one interactor benefits while the other suffers (parasitism). Lastly, interactions can occur in which one microbe is unaffected, while the other is exclusively helped (commensalism) or harmed (amensalism). Networks are useful ways to model these forms of ecological relationships between microorganisms.

Typically, the ecological effect of one microorganism on another can be described by the sign of the interaction (eg. positive, negative, or neutral) and the magnitude of the interaction (eg. strong, weak). The bidirectional ecological relationship between microbes can thus be described using a coordinate pair (x, y) on a Cartesian grid (Figure 2.1), where x represents the net effect of microorganism A on microorganism B , and y represents the net effect of microorganism B on microorganism A . As reviewed by Faust[109], this mathematical framework thus analogizes the five familiar ecological interaction mechanisms, wherein microbes exert mutual effects on one another: competition $(-, -)$, mutualism $(+, +)$, parasitism $(+, -)$, commensalism $(+, 0)$, and amensalism $(0, -)$. Each of these network formalisms have interpretable graphical representations, which are shown in Figure 2.1.

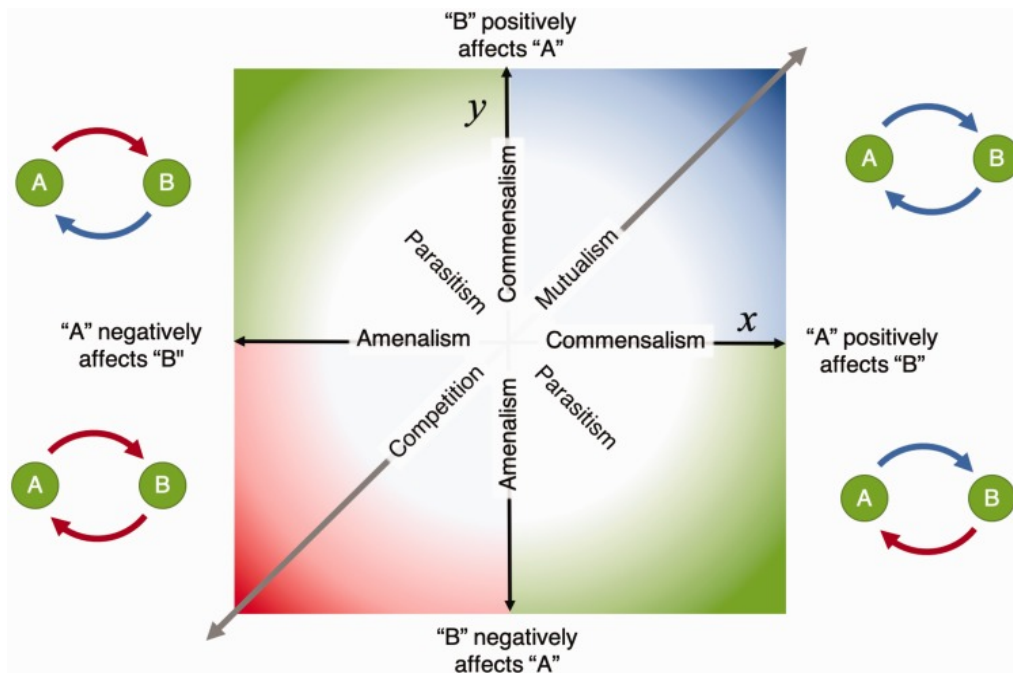


Figure 2.1: The Cartesian plane of pairwise ecological interactions.

Ecological interactions between pairs of microorganisms can be characterized by an (x, y) coordinate pair on a Cartesian plane, where x denotes the effect of microbe A on microbe B, and y characterizes the effect of microbe B on microbe A. The sign of x and y denotes whether the effect of the interaction is positive or negative, while the magnitude denotes the strength of the effect. Classically defined ecological relationships of mutualism $(+,+)$, competition $(-,-)$, and parasitism $(+,-)$ fall into the four quadrants, while amensalism $(-,0)$ and commensalism $(+,-)$, in which one organism is unaffected, lie along the axes. Weighted, undirected ecological networks only capture mutually positive or negative relationships such as competition and mutualism, and are therefore constrained along the diagonal, such that $x = y$.

These bidirectional ecological interactions fit nicely into mathematical framework of networks, allowing further characterization of network models of ecological interactions. In graph and network theory, networks are typified by the edges

they contain. A network is called “weighted”, if we can quantify the strength or magnitude of a given interaction (Figure 2.2A), and “signed” if the weights can take on both positive and negative values (Figure 2.2B). A weighted, signed network is classified as “directed” if the relationships can be described in terms of source and target (or cause and effect) using the aforementioned coordinate pair (x, y) . Directed edges are typically represented by arrows that designate the source and target of an interaction (Figure 2.2C). Undirected networks, however, merely describe mutually positive or negative ecological relationships such as mutualism or competition, but do not delineate the direction of causality for either interactor, rendering commensalism indistinguishable from mutualism, amensalism indistinguishable from competition, and the presence of parasitism largely ambiguous. Only directed networks that are weighted and signed are capable of describing all five forms of ecological interactions mentioned above.

While the concepts of ecological interactions are straightforward in principle, their precise detection of ecological interactions in experimental data remains a significant challenge in the field of microbial network inference [110]. When network inference is the goal of microbial data acquisition, the experimental design depends largely on whether one’s objective is to construct a directed or undirected network. Approaches for identifying interactions in gut microbial ecosystems can thus be broadly classified by their underlying experimental design. Cross-sectional microbiome data,

which consists of static snapshots of multiple individuals, can be used to detect or predict interactions, while longitudinal data, which involves repeated time-series measurements of one or more individuals, can be used to clarify the ecological mechanisms of such interactions. Broadly speaking, undirected, signed, and weighted microbial interaction networks can be inferred from cross-sectional sampling of metagenomic data, while directed network inference requires the collection of time-series, or longitudinal data [111, 112].

2.3 Cross-Sectional Methods for Detecting Microbial Interactions

Undirected, weighted interaction networks, which may indicate positive or negative associations but not causal relationships, can be constructed using a variety of methods. Broadly speaking, these statistical methods are be classified as parametric if they assume adherence to a particular statistical model, or non-parametric if they do not. The simplest and most familiar way to quantify the strength of interactions is using their correlation, and in most data analysis pipelines, the standard parametric statistic for calculating correlation is covariance. While the computation of covariance itself is straightforward under normal conditions, doing so for microbiome data remains a substantial challenge. Being compositional, data describing the proportions of species in microbial surveys are normalized such that the total abundance of a sample sums to a constant value. The result of this normalization is that an increase in the proportion of

Firmicutes, all else held equal, is inherently coupled with an apparent decrease in the proportion all other phyla, resulting in spurious negative correlations by many common statistical methods [113]. Biases stemming from compositional effects and sparsity plague standard covariance metrics in these cases, motivating the development of alternative statistical methods for estimating covariance in compositional microbiome data [105, 106]. Since the overall microbial load of these populations cannot be measured directly, special statistical methods for estimating the covariance matrix must be used instead of calculating it directly.

A handful of parametric statistical methods for inferring the true covariance matrix in compositional microbiome data have been implemented as software programs, many of which are summarized in Table 2.1. One of the earliest such methods, SparCC [107], estimates the covariance using an iterative bootstrap selection procedure. Although designed to assume high data sparsity, it underperforms when sample diversity is high, and is prone to false negatives when the true number of true microbial interactions is large. Another approach, SPIEC-EASI [114], uses Aitchison's centered log-ratio (CLR), a common compositional data transform, and performs well on datasets with high diversity. SPIEC-EASI has been expanded to allow for cross-domain associations, such as between bacteria and fungi [115]. Other parametric methods use a regression method known as LASSO, a statistical technique that in this context,

penalizes excessively complex microbial interaction networks. In particular, CCLasso [116] and REBACCA [117] show improved covariance estimation performance using this technique. Another LASSO-based method, BAnOCC [118], uses a Bayesian approach to estimate covariance, and thus has the benefit of providing uncertainty quantification for network predictions. Finally, MPLasso [119] attempts to incorporate biological prior knowledge into its LASSO approach, by performing automated text-mining of PubMed abstracts to improve performance. Approaches that leverage the existing literature of microbe-microbe interactions are likely to be decisive in the construction and validation of microbial interaction networks.

Although these parametric methods for estimating covariance benefit from interpretability and utility for downstream data analysis, they, like direct covariance calculation methods, are only reliable for detecting linear dependencies between microbes [116]. Other non-parametric strategies of identifying non-linear interactions between microbes have been proposed. For example, mutual information measures such as MIC [120] rely on a measure of association borrowed from information theory to predict functional relationships between variables. A major advantage of MIC has over parametric methods is its ability to capture a broad range of non-linear microbial relationships. It does so by measuring the degree of noise present in potential interactions, rather than the shape of the interaction function itself. This approach

Table 2.1: Summary of statistical tools for inference of weighted, undirected microbial interaction networks from cross-sectional metagenomic data

	SparCC	CCLasso	BAnOCC	MPLasso	SPiEC-EASI	CoNet	MIC	MENA
Full name	Sparse correlations for compositional data	Correlation inference for compositional data through LASSO	Bayesian analysis of compositional covariance	Microbial prior LASSO	Sparse inverse covariance estimation for ecological association inference	Co-occurrence network analysis / renormalization and bootstrap	Mutual information coefficient	Molecular ecological network analysis / Random matrix theory
Implementation	Python	R	R	R	R	Cytoscape plugin	Java	Webapp
Description	Performs empirical covariance estimation using a bootstrapping procedure	Infers the correlation network for latent variables of compositional data using least squares L1 penalty	Estimates a sparse precision matrix through a LASSO prior and generates a posterior distribution by MCMC sampling	Integrates graphical LASSO of microbial co-occurrences with associations obtained from automated text mining of scientific literature	Infers underlying graphical model from conditional independence, using sparse neighborhood and inverse covariance selection	Uses an ensemble method based on multiple similarity measures in combination with generalized boosted linear models	A non-parametric method that detects various noisy distributions by data partitioning	Constructs ecological association networks through random matrix theory
Pros	<ul style="list-style-type: none"> Performs well under low diversity and high sparsity Performs well on sparsity 	<ul style="list-style-type: none"> Covariance matrix is guaranteed positive semi-definite 	<ul style="list-style-type: none"> Low FPR Estimates uncertainty of estimates 	<ul style="list-style-type: none"> Incorporates prior knowledge from literature 	<ul style="list-style-type: none"> Performs well on large feature numbers (OTUs) 	<ul style="list-style-type: none"> Low FPR Easy to use for non-programmers 	<ul style="list-style-type: none"> Detects non-linear relationships Insensitive to rarefaction 	<ul style="list-style-type: none"> Low FPR Robust to noise
Cons	<ul style="list-style-type: none"> Covariance matrix not positive semi-definite Biased towards positive correlations Only measures linear relationships 	<ul style="list-style-type: none"> High FPR Only measures linear relationships 	<ul style="list-style-type: none"> Biased towards positive correlations Only measures linear relationships 	<ul style="list-style-type: none"> Biased towards known interactions Only measures linear relationships 	<ul style="list-style-type: none"> Biased towards positive correlations 	<ul style="list-style-type: none"> Sensitive to distribution type 	<ul style="list-style-type: none"> Requires large sample number Does not quantify direction of associations 	<ul style="list-style-type: none"> Sensitive to distribution type No user-defined P-value
References	Friedman and Alm ³⁰	Fang and Deng, ³⁸ H.	Schwager and Huttenhower ⁴⁰	Lo and Marculescu ⁴¹	Kurtz and Bonneau ³⁶	Faust and Huttenhower ⁴⁷	Reshef and Sabeti ⁴²	Deng and Zhou ⁴⁶

Note: The table contains a brief description of the tool, the statistical methods underlying it, and some of the strengths and weaknesses of each approach.

revealed that co-exclusionary relationships represent a highly common association type between microbiota, as well as that many of the strongest non-linear relationships were dependent on external factors such as diet, sex, and collection method. Another non-parametric approach, LSA [121], was developed for identifying interactions among marine bacterioplankton, and is also capable of detecting non-linear dependencies between microbes. Although LSA was designed with a focus on time-series data, this method can generate undirected networks from static data if the time-delay parameter is set to zero. An expansion of this method, eLSA [122], was developed for time-series with replicates, as well as for approximating the statistical significance of its inferred relationships [123]. Another non-parametric tool, MENA [124], was developed for characterizing microbial interactions in soil, and is highly robust to noise [112]. MENA is based on methods from random matrix theory, a set of statistical tools borrowed from physics. Although non-parametric methods are capable of capturing noisy and highly non-linear microbial interactions, they do not necessarily indicate whether microorganisms are positively or negatively associated, and therefore may only capture the magnitude of an interaction. This means that while non-parametric methods may predict a broader range of microbial interactions, the nature of these interactions may be ambiguous or difficult to model.

While both parametric and non-parametric methods are capable of capturing broad ranges of ecological relationships, it is unlikely that any one method is general enough to detect them all, or even detect them with similar efficiency [120]. Researchers must decide on which tool is right for their application, or otherwise rely on a combination of methods. A comprehensive meta-analysis of microbiome correlation metrics by Weiss [112] showed that precision of network inference could be dramatically improved by the use of ensembles. Because different statistical tools make different mistakes, relying on consensus across tools may be a powerful way to improve the accuracy of microbiome interaction networks. For example, CoNet [125] combines multiple parametric and non-parametric similarity measures with generalized boosted linear models to predict microbial network interactions. As a result, it has a significantly lower false positive rate than other cross-sectional methods [112]. CoNet has been used to identify interactions between species in the skin microbiome [126], as well as among pathobionts associated with cancer cachexia [127]. Since it is integrated with the network visualization software Cytoscape, users quickly and easily construct and analyze microbial interaction networks

Despite the number of correlation methods available for inferring microbial interaction networks, there remains considerable room for improvement. The meta-analysis of eight such methods by Weiss [112] revealed an astonishing degree of

variation in the sensitivity and precision of these tools. On average, methods shared less than a third of predicted interactions. Furthermore, they showed that precision depended greatly on the sequencing technology used, and that normalization choices have a significant impact on edge detection. Next, Weiss simulated several linear ecological relationships to compare tool detection performance. While nearly all methods surveyed were able to detect mutualism and commensalism, amensalism and parasitism were nearly undetectable by the majority of tools. Most concerning however, was that precision was near zero for datasets with more than 50% sparsity. This suggests that abundance filtering is an important first step for detecting correlations between microbes, and that network inference for low-abundance OTUs remains an important area for improvement. Overall, they found that LSA, MIC, and SparCC were the most robust to distribution shape. SparCC performed best in cases where compositionality was high, while LSA did well at capturing both linear and non-linear ecological relationships, even under sparse conditions. LASSO-based methods were not tested, indicating that further work must be done to understand how the LASSO technique performs relative to previously described methods. However, the analysis did demonstrate that ensemble methods show promise in improving precision, particularly for highly sparse datasets. Until sufficiently general methods can be developed,

combining methods with complementary strengths appears to be the best way to improve edge detection for microbial network inference.

2.4 Longitudinal Methods for Detecting Microbial Interactions

Fundamentally, all inference procedures for directed networks are concerned with determining causal relationships between discrete entities. Directed networks typically require longitudinal measurements to infer the source and target of a pairwise microbial relationship. Such methods are expected to be highly important for advancing dynamic models of microbial interactions [128], and may lead to highly precise manipulations of microbial ecosystems. Generally speaking, longitudinal data provide significantly more information on the dynamics of microbial interaction networks than cross-sectional methods. This is because even if the same number of replicates for a cross-sectional study are taken as timepoints for a longitudinal one, the ability to chronologically arrange discrete snapshots of microbial data allows the relational ordering of ecological events. Temporal tracking of blooms and busts in microbial populations thus facilitates the inference of directed microbial interaction networks, as time-delays can be used to indicate causal relationships. A summary of statistical tools for longitudinal network inference can be found in Table 2.2.

Table 2.2: Summary of statistical tools for inference of directed microbial interaction networks from longitudinal metagenomic data.

	LSA	LIMITS	MetaMIS	MC-TIMME	MDSINE	TIME
Full name	Local similarity analysis	Learning Interactions from microbial time series	Metagenomic microbial interaction simulator	Microbial counts and trajectories in infinite mixture model engine	Microbial dynamical systems inference engine	Temporal insights into microbial ecology
Implementation	R	Mathematica	Desktop app	Matlab	Matlab	Web app
Description	Detects complex, non-linear dependence associations between species and environmental factors without data reduction	Combines sparse linear regression with bootstrapping aggregation to infer a discrete-time Lotka–Volterra model	Uses an abundance-ranking strategy paired with partial least square regression to infer a discrete-time Lotka–Volterra model	Models time-varying counts of microbial taxa using an exponential relation process, coupled with adaptive Bayesian techniques	Provides a comprehensive toolbox for dynamical systems analysis of microbiota time-series to fit generalized Lotka–Volterra differential equations	Provides a workflows for time-series metagenomic data, including network inference using Granger LASSO causality
Pros	<ul style="list-style-type: none"> • Can be used on cross-sectional data 	<ul style="list-style-type: none"> • Controls for error due to experimental uncertainty • Bootstrapping reduces compositional effects 	<ul style="list-style-type: none"> • User-friendly interface and visualization • Performs well on rare species 	<ul style="list-style-type: none"> • Allows design of longitudinal experiments • OTU-binning improves estimations 	<ul style="list-style-type: none"> • Multiple algorithms implemented 	<ul style="list-style-type: none"> • Provides user-friendly visualization
Cons	<ul style="list-style-type: none"> • Analysis result is affected by time scale 	<ul style="list-style-type: none"> • Biased towards highly-abundant OTUs 	<ul style="list-style-type: none"> • Biased towards highly-abundant OTUs 	<ul style="list-style-type: none"> • Assumes instantaneous transitions between dynamics 	<ul style="list-style-type: none"> • Requires concentrations, rather than relative abundances 	<ul style="list-style-type: none"> • Does not adequately account for sparsity • Granger causality is prone to high FPR
References	Ruan and Sun ⁴³	Fisher and Mehta ⁵¹	Shaw and Wang ⁵²	Gerber and Bry ⁵³	Bucci and Gerber ⁵⁴	Baksi and Mande ⁵⁵

Note: The table contains a brief description of the tool, how the statistical methods underlying it, and some of the strengths and weaknesses of each approach.

While true causality can only be determined using controlled experimentation [129], mathematical definitions of causality have been applied for predicting causal relationships from time-series data by comparing the histories of related entities [130, 131]. One statistical tool for causal inference is Granger causality [132], which was originally developed for economics but now been used extensively in neuroscience[129]. For a pair of time series, X and Y , we say that X “Granger causes” Y if the histories X and Y together predict the current value of Y better than the history of Y alone. Popular in part due to its computational simplicity [133], this definition of causality has also proven helpful for inference of causal relationships in microbiome studies. TIME [134] is a toolkit that provides a suite of analysis and visualization tools for microbial ecology analysis, and relies on a Granger-LASSO model to identify causal relationships.

Another way to construct directed microbial interaction networks from longitudinal data is to use goodness of fit to a defined model as evidence of causality. Perhaps the most straightforward model for time-dependent ecological modeling of microbiomes relies on generalized Lotka-Volterra (gLV) equations, which are commonly used to describe predator-prey interactions in ecology. This simple mathematical system was first developed by Lotka to describe autocatalytic chemical reactions [135], but was also derived independently by Volterra in early mathematical biology [136]. Fundamentally, gLV defines the growth rate of a given organism as function of the

abundances of all other organisms in a given ecosystem, producing a set of ordinary differential equations. Although these equations are best known for modeling macro-ecological systems, evidence suggests this framework may be applicable to microbiology as well. Here the network inference procedure involves deterministically estimating the interaction terms that determine the dynamics of pairwise ecological relationships, if any. One of the earliest gLV approaches to network inference of microbial time-series data came from a group using multilinear regression to identify interactions in cheese-making microbial communities [137]. This inspired similar work to predict the gLV interaction terms of *Clostridium difficile* infection using regularized regression [138]. Since then, a variety of gLV-based software tools been developed for general applications to microbial network inference. For example, LIMITS [139] uses sparse linear regression to determine the interaction coefficients for a gLV model of microbial interaction dynamics. To overcome the compositional effects associated with relative abundances, LIMITS uses a stepwise approach, iteratively adding edges that produce the lowest error. Another software platform, MetaMIS [140], relies on the use of a partial least square regression to identify the interaction terms, and is implemented as a graphical user interface with tools for network visualization. To maximize the identification of conserved interaction networks, MetaMIS uses an abundance-ranking strategy that prioritizes the identification of interactions between highly abundant microbes, although this strategy

may overlook some novel interactions as a result. While gLV methods represent a powerful and ecologically relevant model, there are drawbacks to their use in microbiome network inference. The microbiome is subject to immigration of new species, spatial variability, and heterogeneity, characteristics that are not necessarily well modeled by gLV dynamics [141]. Additionally, microbiota are understood to interact through a set of complex mechanisms, such as metabolic exchange, that may not be well modeled under this paradigm [142]. And given that gLV models are occasionally unable to detect even pairwise interactions [143], the use of this framework may be questionable. Therefore, while the gLV equations are among the most well-characterized mathematical frameworks for modeling ecological interactions, network inference techniques based on this framework have may not be able to fully capture the intricacies of gut microbiome dynamics. Indeed, there is a demonstrable need for network inference techniques that are able to reliably capture ecological relationships with more complex interaction mechanisms.

One promising area of development for network inference on time-series microbiome data involves the use probabilistic time-series models. Like gLV, these tools are aimed at generating forecasting future behavior using a defined model but are typically better able to handle uncertainty. Such methods have been used extensively in genomics modeling [144-146], and have recently been expanded and adapted for the

purpose of network inference on microbiome data. One such method is MC-TIMME [147], which uses a continuous-time dynamical model and a non-parametric Bayesian technique to identify interaction network. This tool performs OTU-level binning based on similarities in their temporal profiles, which allows improved estimations of the parameters regulating the dynamics of microbial interaction networks. Another probabilistic method, MDSINE [148], constitutes a comprehensive toolkit for dynamical systems inference, as multiple options for inference techniques provided. However, since MDSINE was designed for concentrations rather than relative abundances, it may give erroneous results in situations where overall microbial biomass does not remain constant, or cannot be otherwise measured. Many other approaches to probabilistic time-series modeling of microbial interactions use Dirichlet multinomial mixtures [149, 150], which is a popular distribution for microbiome statistical analysis because it assumes that variables sum to certain value, and is therefore a natural way to model the compositional, relative abundances of microbes. Dynamic linear models, commonly used in commercial forecasting and control engineering [151], may also be useful for modeling microbiome dynamics. While both probabilistic time-series models and dynamic linear models are widely used techniques in other fields, they have only recently been applied to models of microbial dynamics. Hence, little information exists about the relative strengths of these approaches in the context of microbial modeling,

demonstrating a need for analyses benchmarking the comparative performance of these tools against both synthetic and validated microbiome survey data. Additionally, the benefits of ensembles of time-series inference techniques should be evaluated, in light of the improvement Weiss [112] achieved by combining correlation measures with complementary strengths. As there is significant variation in the statistical approaches used to model longitudinal data, it is likely that ensemble approaches will similarly boost performance in the context of longitudinal models as well.

Indeed, dynamical systems and probabilistic models provide a powerful framework for inference of directed microbial interaction networks from longitudinal data [152-154]. When applied correctly, such methods have significant advantages over cross-sectional or correlation-based approaches [155]. By precisely defining the rules and parameters governing microbial interactions, inferred dynamical models are able to characterize the stability of an ecosystem in response to perturbation, and predict future behavior by simulating hypothetical ecosystems [131, 156, 157]. Yet design of longitudinal microbiome studies is not without pitfalls: researchers must strike balance between the cost of data collection and the duration and frequency of sampling. This requires insight into the time scales in which microbial ecosystems fluctuate [157]. Relative to macro-ecological systems, the time scale of microbial interactions is expected to be small [158], making acquisition of usable time-series data challenging or

uninformative [159]. Typically, longitudinal *ex vivo* studies sample participants on the order of days, while both *in vivo* and *in vitro* studies demonstrate that microbial dynamics likely operate on the scale of hours [160-162]. Furthermore, given that metagenomic measurements are discrete and likely to capture steady-state behavior only [13], it may not be possible to capture microbiomes during state-transition, complicating network inference procedures on longitudinal data. While more development of statistical methods for time-series microbiome data is needed, optimization of the frequency and duration of sampling for longitudinal microbiome studies may be equally important.

2.5 The Expanded Universe of Microbial Interactions

While the inference of microbe-microbe interaction networks is a crucial step towards understanding the dynamics of the human microbiome, it somewhat abstracts the true mechanisms by which these interactions occur. Like macro-ecological systems, relationships between microbial species are largely dictated by their food source. While some bacteria, such as ciliates, consume other bacteria, most microbes consume metabolic byproducts excreted by their neighbors, giving rise to some of the ecological dynamics described in Figure 2.1. Mutualistic relationships such cross-feeding, or competition for metabolic resources are largely driven by the import and export of these compounds [163]. Microbial interactions are mediated by metabolites and

macromolecules that are broadly derived from three sources: (1) other endogenous microbiota, (2) endogenous host cells, and (3) exogenous environmental exposures, including dietary intake [164, 165] and chemical exposure [166]. Graphical network representations of metabolite-mediated, host-mediated, and environment-mediated interactions are illustrated in Figure 2.2D-F.

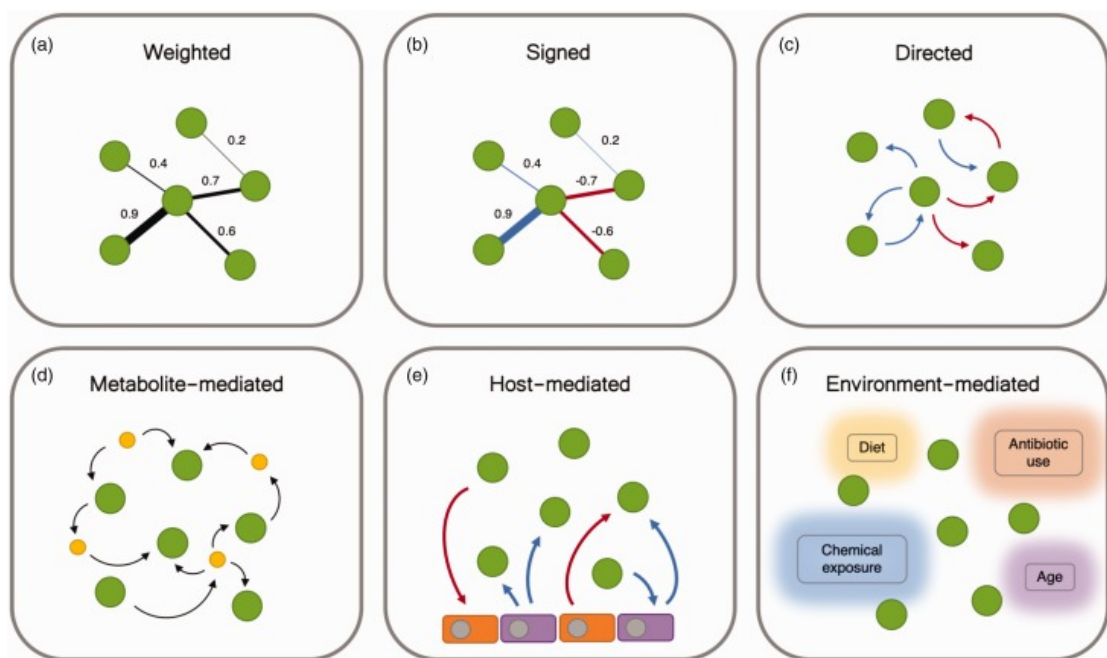


Figure 2.2: Network abstractions of the microbial interactome.

- (A) Weighted networks characterize the strength of an interaction, but do not indicate whether the interaction is mutually positive or negative. Interactions characterized by non-linear relationships take this form.**
- (B) Signed microbial interaction networks denote both the strength and direction of correlations between microbes, but do not indicate a causal relationship. Such networks are typically produced from cross-sectional data.**
- (C) Directed microbial networks characterize source and target of an interaction, indicating a causal relationship. Such networks can be described using the**

ecological terms in Figure 2.1, and are typically produced from longitudinal (time-series) data.

- (D) Interactions between microorganisms are largely mediated by metabolites and macromolecules, which may be consumed or produced as a food source or waste.
- (E) Host cells play an important role in the microbial ecosystem. Host cells may affect the growth of microbes by secreting metabolites or antibiotics. Microbes break down and produce metabolites and macromolecules like short-chain fatty acids, which act as an energy source, and promote the differentiation of host cells in turn.
- (F) Environment-mediated microbial interaction networks contain context-dependent edges. These variables may conditionally alter the topology and dynamics of the microbial interaction networks.

The best approach to simultaneous measurement of the gut microbiome and the gut metabolic state is not obvious. The metabolic activity of microbiota can be assessed indirectly using shotgun metagenomic sequencing, by identifying marker genes associated with metabolic functions [167]. A variety of software tools have been developed towards this end. For example, MicrobiomeAnalyst [168] allows metabolic network visualization from metagenomic sequencing. Another network-based tool, PMRT [169], was developed to predict community metabolic functions from metagenomic data, and HUMAnN [170] can also be used to infer the functional and metabolic potential of microbial metagenomes. If only 16S data is available, tools such as PICRUST [171], Tax4Fun [172], and PiPhillin [173] can be used in combination with KEGG metabolic gene annotations to infer the metabolic state of a community. These

approaches are limited, however, as meta-omics analyses have shown that marker genes for metabolism identified in metagenomic data may not be expressed [174, 175], and genes inferred from 16S sequencing may not be present at all. Of course, the metabolic state of the gut can be measured directly, using methods such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), which can then be integrated with microbiome survey data. Such multi-omics approaches, such as the one employed by Perez-Cobas [176] have potential to dramatically improve our understanding of metabolite-mediated microbial interactions. Methods for integrating such datasets, however, remain in their infancy. The aforementioned PRMT was extended to include integrated analysis of metagenomic and metabolomic data, using a tool called MIMOSA [177]. Otherwise, multivariate correlation methods such as two-way orthogonal partial least squares (O2-PLS) [178], canonical correlation analysis (CCA) [179], and co-inertia analysis (CIA) [180], may be applied to identify interactions between and within omics datasets [180, 181]. Identification of microbiome-metabolite interactions may also be guided by the literature. For example, NJS16 [182], perhaps one of the most comprehensive maps of microbial-metabolite-host interaction networks, was constructed using an exhaustive review of experimental data and existing biological knowledge from the literature. As meta-omics experimental approaches improve, statistical tools for inferring microbiome-metabolite interaction networks must also be

developed, integrated with prior knowledge, and tailored to the unique statistical properties of microbiome data.

Incorporation of host-microbiome dynamics into microbial interaction networks represents an even greater challenge. However, this will be an important step for the development of a comprehensive models of gut ecosystem dynamics. Microbes ferment short chain fatty acids, such as butyrate, which play important roles in host cell function, both as an energy source, and by regulating host gene expression and inflammatory response [183, 184]. Gut microbiota may also produce toxic metabolic byproducts, such as reactive oxygen species, that impair host cell function and promote disease [185]. Conversely, host cells affect the metabolism of resident microorganisms. For example, goblet cells secrete mucin, while hepatocytes mediate glycine- and taurine-conjugated bile acid export [182]. Microbe-interacting host cells are therefore influential components of the gut ecosystem, and models of microbial interaction networks will become more comprehensive with their inclusion. Although the metabolic dynamics of host cells has been reconstructed extensively [186, 187] in a number of model organisms [188], simultaneous measurement of microbial composition and host gene expression or host metabolism in humans is a challenge to perform non-invasively. One potential approach employed by Knight et al. [189] involved extraction of RNA from infant stool samples containing both microbial populations and exfoliated epithelial cells. An *in vitro*

approach using “artificial gut” microfluidics-based human-microbial co-culture systems such as HuMiX [190] may also be a promising avenue for exploring host-microbiome interactions [191]. Alternatively, model organisms may be used. In a review of this topic, Kostic [191] suggests the use of bobtail squid, *Drosophila*, zebrafish, and mice as alternatives to human models for the interrogation of host-microbiome dynamics. However, the clinical relevance of host-microbiome interaction networks developed using non-human experimental models remains circumspect. Even as technology for simultaneous measurement of human and microbial cellular populations improves, statistical approaches must continue to be developed for integrating omics datasets.

Lastly, it is well known that environmental factors shape microbial communities. Microbiome composition varies significantly according to age [192], geography [192, 193], ethnicity [193], diet [194], social networks [195], and chemical exposure [166]. Consequently, inferred microbial interaction networks are likely to be significantly influenced by their environmental contexts. Diet in particular is an important environmental variable, as it strongly influences the metabolic environment of the gut microbiome [196]. Identifying interactions between environmental characteristics and microbiome composition can be done using aforementioned multivariate approaches like CCA, O2-PLS, and CIA [197]. Understanding how and when environmental factors will conditionally influence the topology of microbial

interaction networks will be necessary not only to control for these factors, but also to understand the degree to which microbiome dynamics are context-dependent.

3 The Cancer Microbiome Atlas: A Pan-Cancer Comparative Analysis to Distinguish Organ-Associated Microbiota from Contaminants¹

3.1 Introduction

The human body supports an ecosystem of 10-100 trillion microorganisms [14, 198], representing 500-1000 unique species per individual [199, 200]. Perturbations to this ecosystem, termed dysbiosis, can impact human health: microbial alterations have been implicated in a variety of health conditions, including obesity, diabetes, inflammatory bowel disease, cancer, and other diseases [201-204]. While public microbiome projects such as the Human Microbiome Project (HMP) and MetaHIT have helped bring tremendous insights into the diversity and function of human flora, these databases are dominated by tissue swabs and stool samples that do not necessarily reflect the microbial composition of internal organs [205, 206]. Collection of clinical biopsies specifically dedicated to microbial profiling remains difficult despite many disease-related host-microbe interactions occurring at the epithelium of internal body sites.

¹ This chapter is exactly reproduced from a research article of the same name authored by A.B. Dohlman, D.A. Mendoza, S. Ding, M. Gao, H. Dressman, I.D. Iliev, S.M. Lipkin, and X. Shen, published in *Cell Host & Microbe*.

Next-generation sequencing (NGS) is frequently used to profile biopsied human tissue samples at a broad range of body sites and disease states, and these sequencing datasets contain a significant number of sequencing reads of microbial origin [62, 71, 72]. Large sequencing projects can thus be mined to promote understanding of host-microbe interactions in both healthy and diseased human tissue. To that end, the bioinformatics tool PathSeq [71] was used to identify enrichment of *Fusobacterium nucleatum* in TCGA colorectal cancer (CRC) tumors [39, 43]. Since then, dozens of research articles explored the role of *F. nucleatum* in tumorigenesis, finding associations with stage, survival, metastasis, and even drug response [28, 207, 208]. More broadly, sequencing data from TCGA has been used *ad hoc* to screen for viral and bacterial presence in stomach adenoma [209] and cervical cancer [210] specifically, as well as viromes [211] and bacteriomes [72]. Recently, analysis of TCGA sequencing data has been used to demonstrate the potential for bloodborne microbial DNA to diagnose certain cancers [62]. Given that even low-biomass tumors contain tissue-specific microbiomes [35], analysis of microbial DNA and RNA in TCGA sequencing data has great potential for diagnostic applications, as well as for exploring host-microbe interactions along molecular and clinical axes.

However, few actionable microbiota targets like *F. nucleatum* have emerged from such analyses. When examining a subset of TCGA sequencing data, previous analyses

[72] found that microbial reads from a number of species were the result of contamination, and that distinguishing contamination from tissue-embedded microbes remained an outstanding challenge for use of this dataset. Indeed, while concerns over contamination are less pressing for samples with high microbial biomass such as stool or swabs, microbiome studies on low-biomass samples suffer from contamination during sample collection and DNA extraction, which can originate from the laboratory environment, including from nucleic acid extraction kits [74-76]. Thus controlling for contamination in these datasets is a crucial step that must precede downstream analyses of host-microbe interactions. Samples for multi-institutional projects are acquired, processed, and sequenced at different sites, each of which may introduce its own contaminants that influence the extracted microbial profiles, impede reproducibility, and complicate discovery of microbial biomarkers. Thus, a number of strategies have been deployed to identify contamination in TCGA sequencing data, through examination of batch effects, sample analyte concentrations, and through manual curation [62, 72]. To date, such analyses have never been validated by original TCGA tissue or blood samples, nor have decontaminated TCGA microbiome datasets been made readily available.

Sequencing data in TCGA provide a unique opportunity for identifying tissue-specific microbiota since matched tissue and blood samples from various cancer types

are processed and sequenced in parallel using various sequencing platforms at designated centers [212]. Using an unbiased statistical model comparing the prevalence of microbial species in tissue and blood samples, we isolated the tissue-resident microbiome in TCGA sequencing data. We found that species equally prevalent across tissue types and blood samples are mostly artifacts or contaminants that 1) bear unique signatures from the designated TCGA sequencing center and 2) comprise more than half of all detectable microbial sequencing reads in many tissue samples. With gene- and nucleotide- level resolution, our model is also capable of normalizing read counts for “mixed-evidence” cases, in which sequencing reads aligning a given species may come from a combination of endogenous and contaminant microbiota. To validate our approach, we obtained original matched CRC tissue and plasma samples that were previously sequenced by TCGA and performed 16S rRNA amplicon sequencing. This independently confirmed not only the absence of putative contaminants but also that the tissue-resident, computationally decontaminated microbial profiles we extracted from TCGA sequencing data matched the microbial composition of the original tissue samples.

Finally, we ran the vetted decontamination algorithm to establish the TCMA database, which users can access from an interactive website (<https://tcma.pratt.duke.edu>). The database contains curated tissue-resident microbial

profiles for 4,937 sequencing runs on 3,689 unique samples from 1,772 patients representing 5 TCGA projects and 21 anatomic sites with tissue-resident populations. As proof-of-principle, we used TCMA to identify two bacterial coabundance groups in CRC tissue, including species enriched in CRC tumors compared to matched adjacent normal tissue, and species prognostic of patient survival. TCMA enabled a matched microbe-host transcriptomic, proteomic, and epigenetic analysis that identified associations between microbes and host gene expression patterns and pathways. Lastly, by comparing TCMA-curated blood samples of CRC and brain cancer (BC) patients, we identified a bacterial signature associated with colorectal mucosal barrier injury unique to CRC blood samples. Thus, TCMA constitutes a powerful resource for validation and hypothesis generation in future studies of host-microbe interactions relevant to cancer.

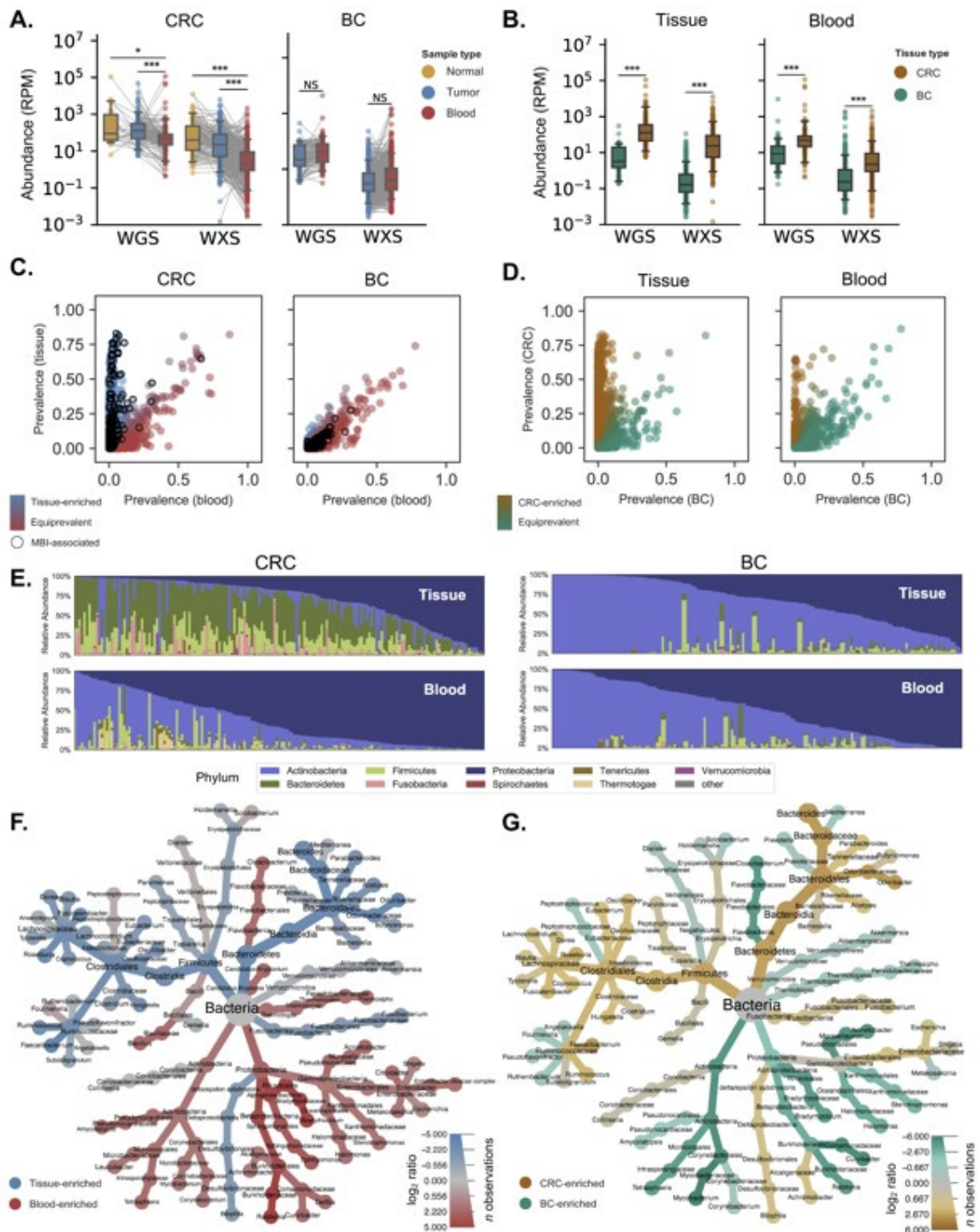


Figure 3.1: WGS and WXS harbor colorectal bacterial reads distinct from blood and brain

- (A) Matched analysis of bacterial sequencing reads per million (RPM) in normal tissue (yellow), tumor tissue (blue), and blood (red) from CRC and BC patients in TCGA. Significance is given by paired, one-sided t-tests.
- (B) Shows abundance data from (A), but compares solid tissue (pooled tumor and normal) with blood samples from BC (green) and CRC (brown) patients. Significance is given by one-sided t-tests.
- (C) Comparison of bacterial species prevalence in WGS data for CRC blood and CRC tissue samples reveals populations of tissue-enriched species (blue) and species that are equiprevalent in blood and tissue (red). Black circles denote species associated with mucosal barrier injury (MBI).
- (D) Comparison of bacterial species prevalence in WGS data for CRC and BC samples reveals populations of CRC-enriched species (brown) and species that are equiprevalent in CRC and BC (green).
- (E) Relative abundance of bacterial phyla in WGS data for tissue (top) and blood (bottom) samples from CRC (left) and BC (right) patients.
- (F) Heat-tree comparing relative abundance of bacteria in WGS data for matched blood-samples (red) vs. tissue samples (blue) (F) and CRC tissue (brown) vs. BC tissue (green) (G).

3.2 Results

3.2.1 WGS and WXS harbor colorectal bacterial reads distinct from blood and brain

To explore the microbial populations of sequenced TCGA tissue, we began by analyzing multi-platform sequencing data for 730 tissue and 555 blood samples from 617 colorectal cancer (CRC; TCGA projects COAD/READ) patients and for 958 tissue and 914 blood samples from 923 brain cancer (BC; TCGA projects GBM/LGG) patients. For

several thousand whole-genome (WGS) and whole-exome (WXS) sequencing experiments, we retrieved raw sequencing data from the TCGA database and extracted and mapped high-quality reads of bacterial origin using PathSeq [71]. We found that microbial reads were more abundant and more diverse in solid tissue than in matched blood samples from CRC patients; in contrast, the abundance and diversity of microbial reads were no greater in BC tissue than matched blood samples (Figure 3.1A, Figure S3.1A). Furthermore, CRC tissue had more abundant and diverse microbiota than BC tissue (Figure 3.1B, Figure S3.1B), consistent with the notion that blood and brain tissue are more sterile than colorectal tissues. Notably, microbial reads were also more abundant and diverse in blood samples from CRC patients than in those of BC patients (Figure 3.1B, Figure S3.1B).

Comparative analysis of microbial reads between CRC tissue and blood samples revealed two distinct groups of bacterial species: those enriched in tissue, and those equally prevalent in tissue and blood (Figure 3.1C, Figure S3.1C). Species were seldom more prevalent in blood than in tissue. Species more prevalent in CRC tissue than in CRC blood included many species known to be associated with mucosal barrier injury (MBI) [213], whereas the group equally present in CRC tissue and blood contained very few such species (Figure 3.1C, Figure S3.1C). By comparison, nearly all species detected in samples from BC patients were equiprevalent in tissue and blood, with few enriched

in tissue (Figure 3.1C, Figure S3.1C). Similar comparative analyses of tissue and blood from CRC and BC patients showed significant populations of disease-enriched species for CRC but few for BC (Figure 3.1D, Figure S3.1D). We then repeated this comparative prevalence analysis using samples from ovarian cancer (OVC; TCGA project OV; Figure S3.1E-F).

The microbial composition of CRC tissue samples was also markedly distinct from that of matched blood samples or BC tissue samples. Among the most dominant phyla in CRC tissue were *Bacteroidetes* and *Firmicutes*, which were relatively absent in blood and brain tissue samples (Figure 3.1E, Figure S3.1G). We next compared the relative abundance of bacterial taxa in CRC tissue vs. matched blood samples (Figure 3.1F) and CRC tissue vs. BC tissues (Figure 3.1G) from different donors. These analyses were largely consistent, with taxa from *Bacteroidetes*, *Firmicutes*, and *Fusobacteria* clades consistently overrepresented in CRC tissue, compared with *Proteobacteria* and *Actinobacteria*, which accounted for a relatively greater fraction of reads in CRC blood samples and BC tissue samples. Genera that were relatively more abundant in CRC blood samples or BC tissue samples compared with CRC tissue samples consistently included *Acinetobacter*, *Mycobacterium*, and *Ralstonia*, among others (Figure 3.1F-G). Metagenomic profiling of TCGA samples using Kraken2 [214] largely recapitulated PathSeq results (Figure S3.1H-J).

Together, these comparative analyses were capable of distinguishing species enriched in CRC tissues from those with similar prevalence across different blood samples and disease types. The analyses confirmed that bacteria in CRC tissues were 1) more diverse and abundant and 2) were enriched for mucosa-related species.

3.2.2 Species equiprevalent in tissue and blood are predominantly contaminants

Besides species enriched in CRC tissues, a significant number of detected species were equally prevalent in blood, CRC tissue, BC tissue, and OV tissue (Figure 3.1C-D, Figure S3.1C-F). While compromised epithelial barrier function may allow the translocation of microorganisms to the bloodstream at low levels [215], we expected that such species would be prevalent at much lower rates in blood than in CRC tissue. To analyze equiprevalent species and determine their origin, we first examined a set of 70 bacterial genera known to be enriched in negative controls of metagenomic sequencing experiments [74]. Overall, genera in this “common contaminant” set were more prevalent in blood samples ($p = 4.45e-10$; Figure 3.2B, Figure 3.2A) than genera not in the list.

Species equiprevalent in CRC tissue and blood were also considerably more genetically and phenotypically diverse than species enriched in CRC with respect to their G-C content, genome size and optimal growth conditions. Conversely, CRC-enriched species were much less tolerant to extreme growth conditions, with optimal

temperature, pH, and NaCl levels more closely resembling those of human homeostasis (Figure 3.2C, Figure S3.2B). Together, these results suggested that the equiprevalent group may contain contaminant species, which have larger genomes, a signature of “generalist” bacteria that must endure more variable and unstable environmental conditions than gut microbiota [216].

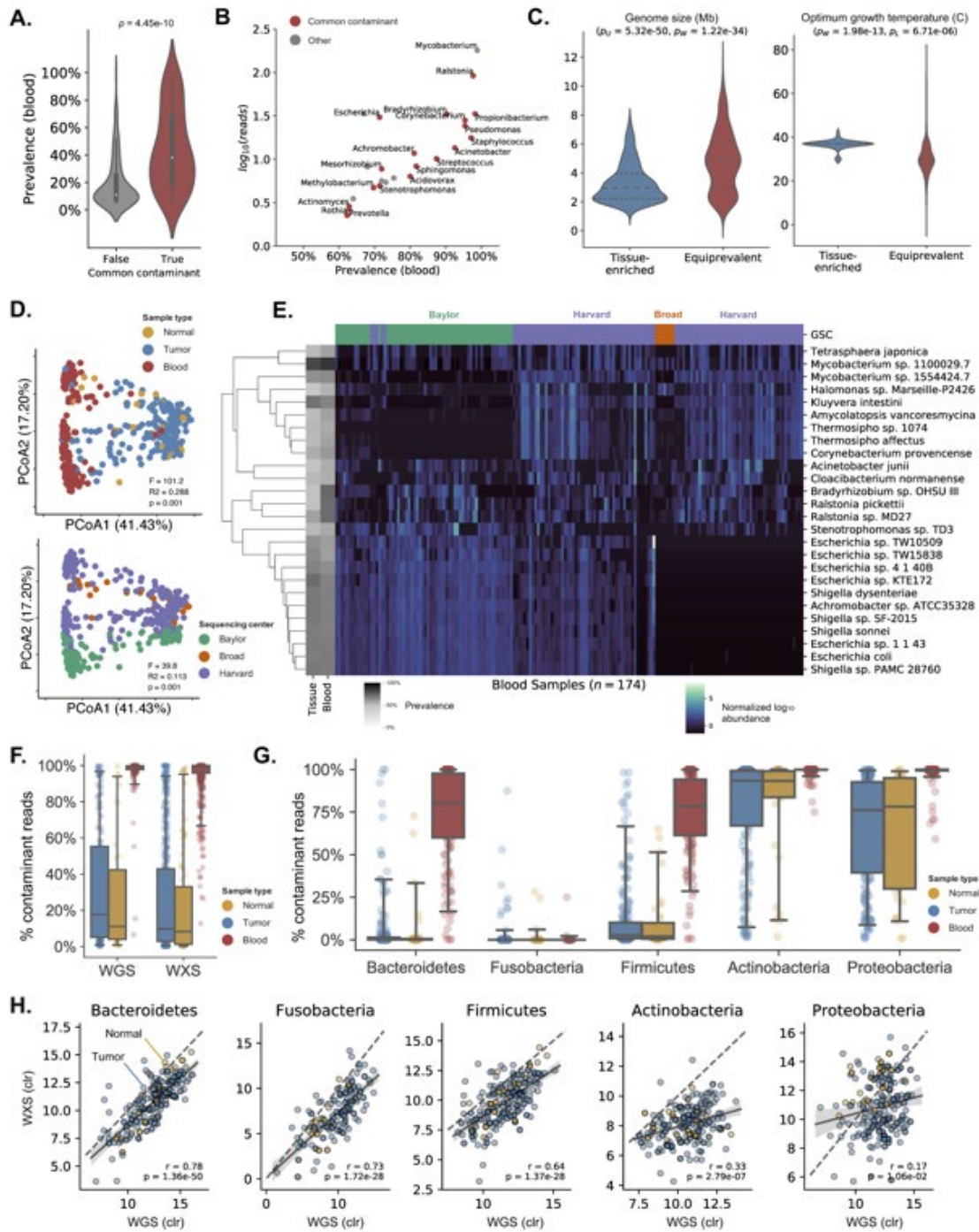


Figure 3.2: Most equiprevalent taxa are common contaminants and associated with particular sequencing centers

- (A) Genera commonly found in negative controls of metagenomic sequencing experiments [74] are highly prevalent in blood samples.
- (B) Prevalence of common contaminants in blood are correlated with absolute abundance.
- (C) Equiprevalent species have differential (pW, Wilcoxon's test) and more variable (pL, Levine's test) genome size and temperature tolerance than tissue-enriched species.
- (D) PCoA of WGS data for CRC samples reveals considerable variation between blood samples and tissue samples along the first axis of variation, and batch effects along the second axis.
- (E) Heatmap clustering of bacterial species' abundance in blood samples demonstrates the presence of center-specific contamination. The left vertical axis shows each species' prevalence (grey).
- (F) Fraction of all bacterial reads that are contamination in normal (yellow), tumor (blue), and blood (red) samples from CRC patients.
- (G) Fraction of bacterial reads that are contamination in WGS data of normal (yellow), tumor (blue), and blood (red) from CRC patients, broken down by the five most prevalent phyla.
- (H) Correlations between centered log ratio (CLR)-transformed relative abundances of WGS and WXS data for the five most prevalent phyla in tissue samples. Phyla contributing the most contaminant reads have the lowest correlation between assays.

3.2.3 Equiprevalent species are associated with particular sequencing centers

Principle coordinate analysis (PcoA) of UniFrac distances between CRC samples demonstrated that the primary axis of variation in TCGA microbiome data could be

attributed to differences between blood samples and tissue samples (41.43%) (Figure 3.2D, Figure S3.2C). Interestingly, the second axis of microbial variation was determined by the sequencing center at which the samples were processed (17.20%), regardless of sample type. All TCGA samples (tissue and blood) were harvested at a tissue source site and then sent to designated genome sequencing centers (Figure S3.2D). While the first PCoA axis captured differences in the presence of tissue-enriched species that are more abundant in CRC tissue than in blood, the second axis captured species found in both tissue and blood samples at similar rates, many of which were associated with sequencing center (Figure 3.2D, Figure S3.2E-F). We then examined the abundance of equiprevalent species in blood samples and found significant clustering according to the sequencing centers at which the samples were processed (Figure 3.2E, Figure S3.2F). To compare, we performed the same analysis on tissue and blood samples from BC patients, which revealed no discernible variation between tissue and blood samples, but rather significant clustering by sequencing center (Figure S3.2G-H).

Therefore, the majority of species equiprevalent in the tissue and blood are not endogenous but are mostly artifacts introduced during processing and profiling at their respective sequencing centers. For ease of description, we will refer to equiprevalent species as “contaminants” and tissue-enriched species as “tissue-resident” for the

remainder of the article. However, the equiprevalent population may still contain biologically relevant species that are detected in both tissue and blood.

3.2.4 A generalizable model for isolating tissue-resident microbiota in TCGA tumor samples

Based on the comparative analyses of prevalence in tissue and blood, we developed a generalizable statistical model to distinguish tissue-resident microbiota from contaminant species across cancer types in TCGA. Of the 1,136 bacterial species detectable in more than 5% of CRC tissue samples, this model classified 769 species as tissue-resident (67.69%) and 367 species as contamination (32.31%). Tissue-resident populations identified by comparing prevalence in tissue and blood were largely consistent with prevalence comparisons of CRC tissue with BC tissue, as well as analogous comparisons made with WXS data (Figure S3.2I). The model was used to perform binary classifications for all observed species. For each sample in the cohort, and at each subsequent taxonomic level, we then used species-level classifications to design a mixture model estimating the fraction of contaminant read counts within a given clade. This method provides a generalizable approach for decomposing observed microbial populations into their tissue-resident and contaminant fractions on a sample-by-sample basis at multiple taxonomic levels.

3.2.5 *Proteobacteria* and *Actinobacteria* contribute the largest fraction of contaminant reads

As expected, the removal of contaminant species resulted in a reduction in the number of bacterial reads in all sample types. Specifically, bacterial species classified as contaminants accounted for a median of 16.27% bacterial reads counts in tissue but varied considerably (Figure 3.2F). Contamination consistently dominated blood samples, with a median of 99.45% detected WGS reads being the result of contamination. The phyla *Proteobacteria* and *Actinobacteria* contributed the greatest fraction of contaminant reads in WGS data, with medians of 76.67% and 80.95% reads found in CRC tissue samples being the result of contamination, respectively (Figure 3.2G). By contrast, only small fractions of *Firmicutes* (1.70%), *Bacteroidetes* (0.02%), and *Fusobacteria* (0.00%) reads were predicted to be the result of contamination (Figure 3.2G). Contamination rates were largely similar for WXS and across sequencing centers (Figure S3.2J-K).

Additionally, correlation between the normalized relative abundances of taxa in matched WGS and WXS samples was predictive of contamination rates. The correlations between *Bacteroidetes*, *Fusobacteria*, and *Firmicutes* abundances in WGS and WXS were consistently high, in contrast to *Actinobacteria* and *Proteobacteria* (Figure 3.2H). For blood samples, the normalized relative abundances of these five domains were wholly uncorrelated between matched WGS and WXS (Figure S3.2L). Overall, these results

show that significant fractions of the bacterial reads in WGS data for CRC tissue and blood samples are the result of contamination from *Actinobacteria* and *Proteobacteria* species.

3.2.6 Detecting tissue-resident and contaminant species with gene-level resolution

For species designated as tissue-resident (e.g., *B. vulgatus*) or as contamination (e.g., *A. junii*), we next explored the extent to which microbial genes could be reliably detected in TCGA sequencing data. Using annotated genomes to search for gene-level assignments, we found that for many such species, sequencing alignments provided coverage of the full microbial genome. As expected, gene prevalence profiles of tissue and blood samples largely recapitulated those of species-level assignments (Figure 3.3A-B, Figure S3.3A-B). For tissue-resident species, the gene prevalence distribution was much lower in blood samples than tissue, while for contaminants, the gene prevalence distribution of blood and tissue samples were nearly identical (Figure 3.3D-E, Figure S3.3C-D); respectively, genome coverage was greater in tissue than blood for tissue-resident species, and identical for contaminant species (Figure 3.3G-H, Figure S3.3E-F). Likewise, genome coverage in tissue samples was nearly equal at Harvard and Baylor for tissue-resident species, but not for contaminant species (Figure S3.3G-H). These results suggested that gene- and nucleotide-level analyses of microbial sequencing reads may be leveraged to help distinguish contamination from tissue-resident populations.

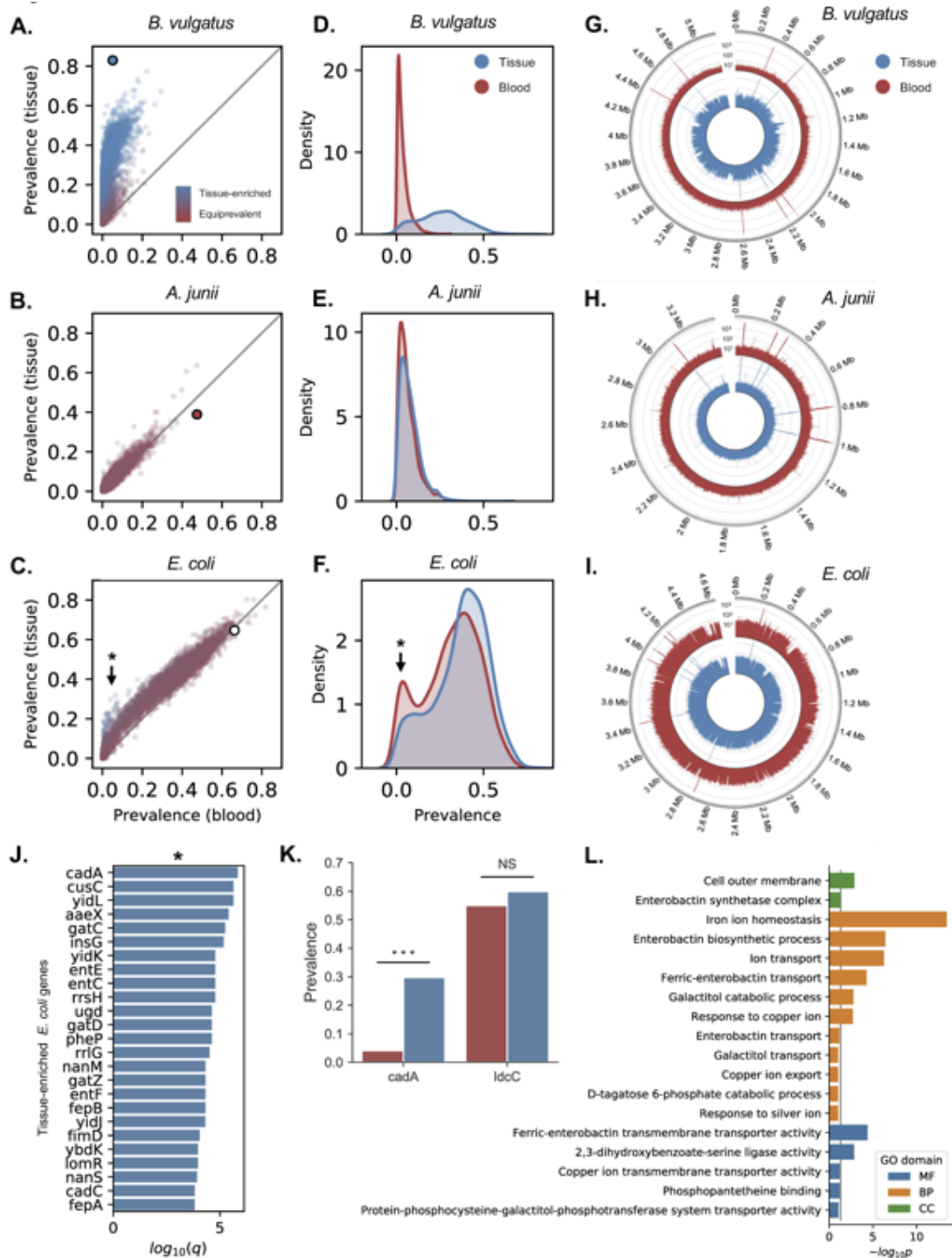


Figure 3.3: Detecting tissue-resident and contaminant species with gene-level resolution

- (A – C) Prevalence of genes belonging to *B. vulgatus* (A; tissue-resident), *A. junii* (B; contaminant), and *E. coli* (C; mixed-evidence) in blood vs. tissue; representative of tissue-resident species. The large dot indicates species-level prevalence.
- (D – F) Kernel-density estimate of gene prevalence in blood (red) and tissue (blue) for *B. vulgatus* (A), *A. junii* (B), and *E. coli* (F).
- (G – I) Coverage of WGS reads aligning to genomes of *B. vulgatus* (G), *A. junii* (H), and *E. coli* (I) in blood (red) and tissue (blue).
- (J) Top 25 *E. coli* genes most significantly enriched in tissue.
- (K) Comparison of *E. coli* genes *cadA* and *ldcC* prevalence in blood (red) and tissue (blue).
- (L) Results of GO pathway analysis of tissue-enriched *E. coli* genes.

3.2.7 Distinguishing tissue-resident *Escherichia* reads from contamination

An outstanding challenge in controlling contamination is the problem of mixed-evidence cases in which detected sequencing reads come from an unknown combination of endogenous and contaminant sources [62, 72]. For example, although *Escherichia coli* is ubiquitous among human microbiomes, species-level *E. coli* reads were present in tissue (64.68%) and blood (66.29%) at nearly equal rates and were strongly associated with sequencing center (Figure 3.2E). We therefore explored whether gene-level read alignments could provide greater resolution and be used to estimate the fraction of sequencing reads resulting from contamination versus endogenous microbiota. To test

this, we mapped microbial sequencing reads from TCGA tissue and blood samples to genes in the annotated *E. coli* genome.

Overall, reads aligning to *E. coli* genes in tissue and blood samples were detected at up to the same rates as species-level *E. coli* alignments (Figure 3.3C) and had similar genome coverage (Figure 3.3I). However, a small number of *E. coli* genes displayed a signature analogous to tissue-resident microbiota in our species-level prevalence analysis (Figure 3.3C). Moreover, we observed bimodality in the blood prevalence of *E. coli* genes (Figure 3.3F), suggesting the presence of distinct tissue-resident and contaminant *E. coli* populations. We identified a set of 119 *E. coli* genes significantly enriched in tissue samples ($q < 0.01$; Figure 3.3J), several of which have credible reasons for being enriched in tissue samples. The top candidate, *cadA* ($q = 4.44\text{E-}9$), is a gene encoding one of two lysine decarboxylases [217, 218] produced by *E. coli*; the other is *ldcC*, which is not enriched in tissue samples ($q = 0.16$) (Figure 3.3K, Figure S3.3J). While *ldcC* encodes a gene that is constitutively expressed, *cadA* transcription is induced under conditions of anaerobic growth at low pH and its gene product displays greater thermostability and acid-tolerance [219]. Additionally, genes in the *pks* island encoding colibactin were significantly more prevalent in tissue than blood, matching previous reports that *E. coli* strains expressing this gene are associated with CRC tissues (Figure S3.3L) [220].

Discrepancies in intraspecies genome content may be explained by adaptive gene loss, an evolutionary mechanism whereby bacteria dispense with genes that are unnecessary for their environmental conditions [221, 222]. Pathway analysis [223] revealed that tissue-enriched *E. coli* genes were significantly associated with processes including iron ion homeostasis, enterobactin biosynthesis, ion transport, ferric-enterobactin transport, and copper ion response ($p < 0.01$) (Figure 3.3L). Iron (Fe^{3+}) and copper (Cu^{2+}) are abundant in the host and can be toxic to *E. coli* in acidic, aerobic conditions; therefore, strains of *E. coli* must tightly regulate intracellular concentrations of these metals, and undergo selection to do so [224, 225]. Given that hypothetically bloodborne *E. coli* would also have to contend with high concentrations of copper and iron, enrichment of these genes and processes in tissue relative to blood suggests that the majority of *E. coli* reads detected in blood samples are not endogenous but rather the result of contamination. However, tissue-enriched genes such as *cadA* and others serve as benchmarks for distinguishing the two.

3.2.8 Tissue-enriched sequencing reads can be identified with nucleotide precision

We then examined microbial sequencing reads at nucleotide-level resolution. Given that gene-level alignments helped resolve mixed-evidence cases, we explored whether bacterial sequence variants such as SNPs could be used in a similar fashion. Variant prevalence across CRC tissue and blood samples largely recapitulated the

results from species- and gene-level profiles (Figure S3.3L-M). Interestingly, we also found populations of apparent tissue-enriched and equiprevalent variants in *E. coli* genomes, suggesting that analyses of sequence variants may prove useful in distinguishing between endogenous and contaminant sequencing reads in mixed-evidence cases.

3.2.9 Decontamination removes sequencing center artifacts

Removing contamination affected all samples, but samples with low bacterial abundance *a priori* were the most affected (Figure 3.4A), consistent with observations that low biomass samples are the most profoundly affected by contamination [74, 76]. Decontamination also regularized the relative abundance profiles of CRC tissue samples, most prominently by the removal of contaminant *Actinobacteria* and *Proteobacteria* reads (Figure 3.4B).

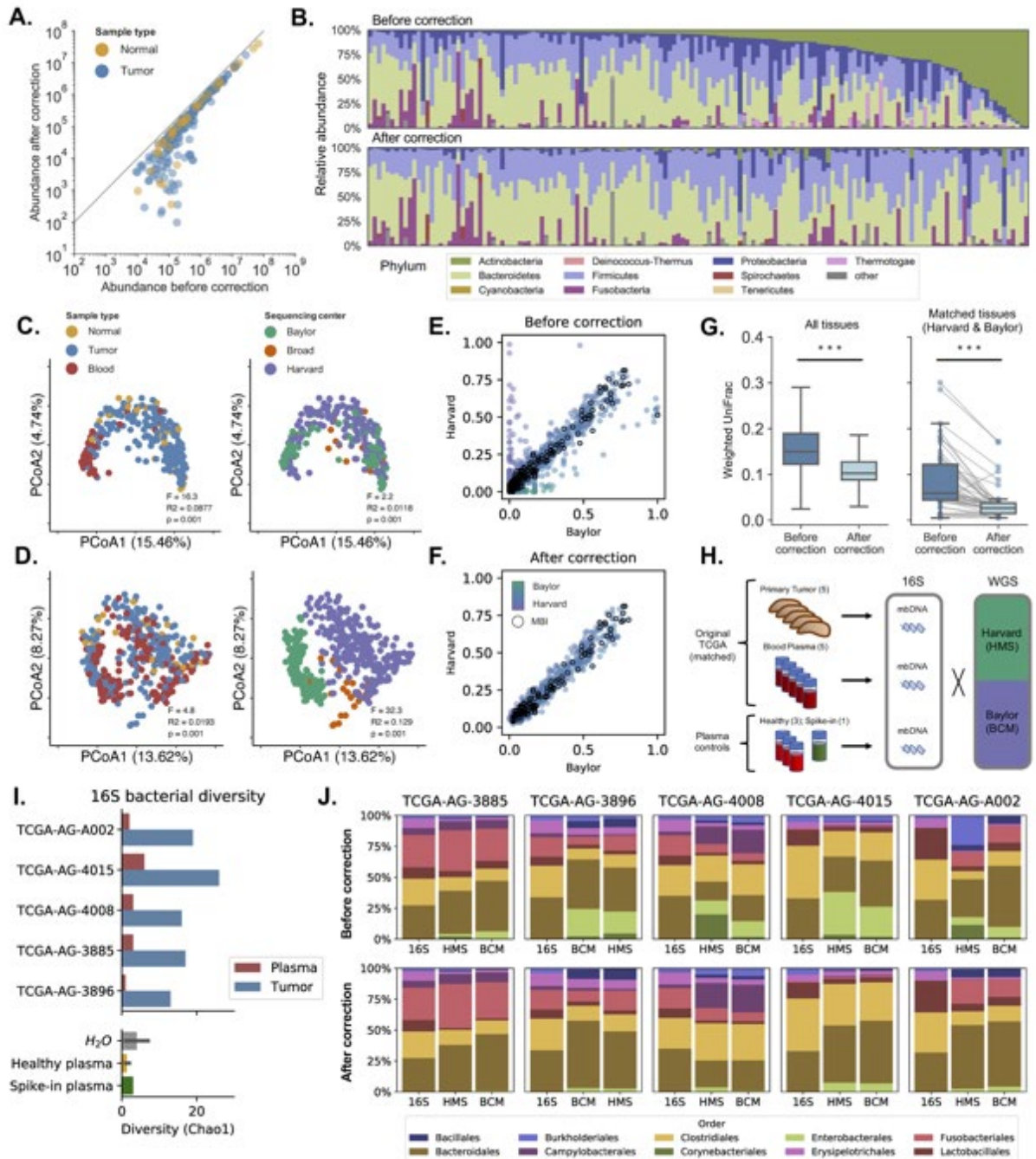


Figure 3.4: Decontamination removes sequencing center artifacts and original TCGA tissue and blood samples validate tissue-resident microbial compositions and equivalent species as contaminants

- (A) **Abundance of WGS bacteria in before and after decontamination. Samples with no reduction in bacterial reads lie along the gray line. Experiments with low microbial biomass a priori are disproportionately affected by decontamination.**
- (B) **Relative abundance of bacterial phyla in tissue samples before and after decontamination, sorted by their a priori abundance of Actinobacteria.**
- (C) **PCoA of the decontaminated, tissue-resident microbial component reveals retention of variation related to sample type, but not sequencing center.**
- (D) **PCoA of the contaminant microbial component reveals retention of variation related to sequencing center, but not sample type.**
- (E – F) **Prevalence of bacterial species in tissue samples sequenced at Baylor vs. Harvard before (E) and after (F) removing contamination.**
- (G) **Comparison of weighted UniFrac distances before and after removing contamination among all tissues (left) and specifically matched tissues sequenced at both Baylor and Harvard (right).**
- (H) **Design of the validation experiment. Data are represented as mean \pm 95% CI.**
- (I) **Bacterial diversity of 16S results from tissue (blue), plasma (red) and controls (bottom panel).**
- (J) **Relative abundances in 16S results for tissue compared with tissue samples sequenced using WGS at Harvard and Baylor, before and after contamination.**

Although our prevalence-based model for decomposing the TCGA microbiome data into tissue-resident and contaminant fractions was naïve to sequencing center, removing contamination also mitigated center-related batch effects. Prior to removing contamination, TCGA microbiome data clustered by both sample type and sequencing center (Figure 3.2D). However, unsupervised clustering of the tissue-resident

component extracted from the original TCGA sequencing data showed no dependency on sequencing center and maintained variation related to sample type (Figure 3.4C). Examining the contamination component, we found the opposite: samples no longer clustered by sample type, but rather organized exclusively according to sequencing center (Figure 3.4D). These results reflected the removal of species that were uniquely prevalent in tissue samples from either Baylor or Harvard (Figure 3.4E-F, Figure S3.4A).

Finally, our algorithm greatly increased the similarity between the microbial populations in patient-matched tissue samples sequenced at both Harvard and Baylor, while maintaining diversity among samples overall (Figure 3.4G). Thus, our prevalence-based model is able to homogenize matched samples sequenced at different centers and mitigate sequencing center artifacts.

3.2.10 Original TCGA tissue and blood samples validate tissue-resident microbial compositions and equivalent species as contaminants

To benchmark our analysis, we obtained five primary CRC tumor samples and matched plasma samples from an original TCGA tissue provider (Table S3.1). These samples were specifically chosen to ensure that each tissue and plasma sample was profiled by WGS at both Baylor and Harvard. For controls, we also procured three plasma samples from healthy individuals and spiked one plasma sample with *E. coli*. We then used 16S amplicon sequencing to validate that the bacterial composition of the

original TCGA samples resembled the decontaminated compositions extracted from TCGA sequencing data for matched tumor samples (Figure 3.4H).

We found that the original TCGA tumor samples contained ample bacterial diversity and read counts (Figure 3.4I, S3.4B) and that their bacterial composition largely recapitulated the decontaminated, tissue-resident microbial population our decontamination model extracted from TCGA WGS data on matched samples (Figure 3.4J, Figure S3.4D). In addition to increasing the similarity between WGS and 16S validation results (Figure S3.4C), decontamination greatly improved the concordance between microbial compositions of matched WGS experiments performed at Harvard and Baylor (Figure 3.4G). Despite detecting a large number of bacterial sequencing reads in tumor samples, bacterial diversity of CRC plasma was not significantly greater than healthy plasma ($p = 0.30$) or water controls ($p = 0.44$). Moreover, the 16S bacterial composition of original TCGA plasma samples was distinct from the bacterial composition of WGS data from the same samples (Figure S3.4E-F), supporting the notion that the majority of bacterial reads detected in TCGA blood samples are contamination introduced during DNA extraction and sequencing, rather than at the time of procurement. These validation results demonstrate that our model accurately identifies and removes contaminants and that computational decontamination produced profiles that represent the true microbial composition of tissue.

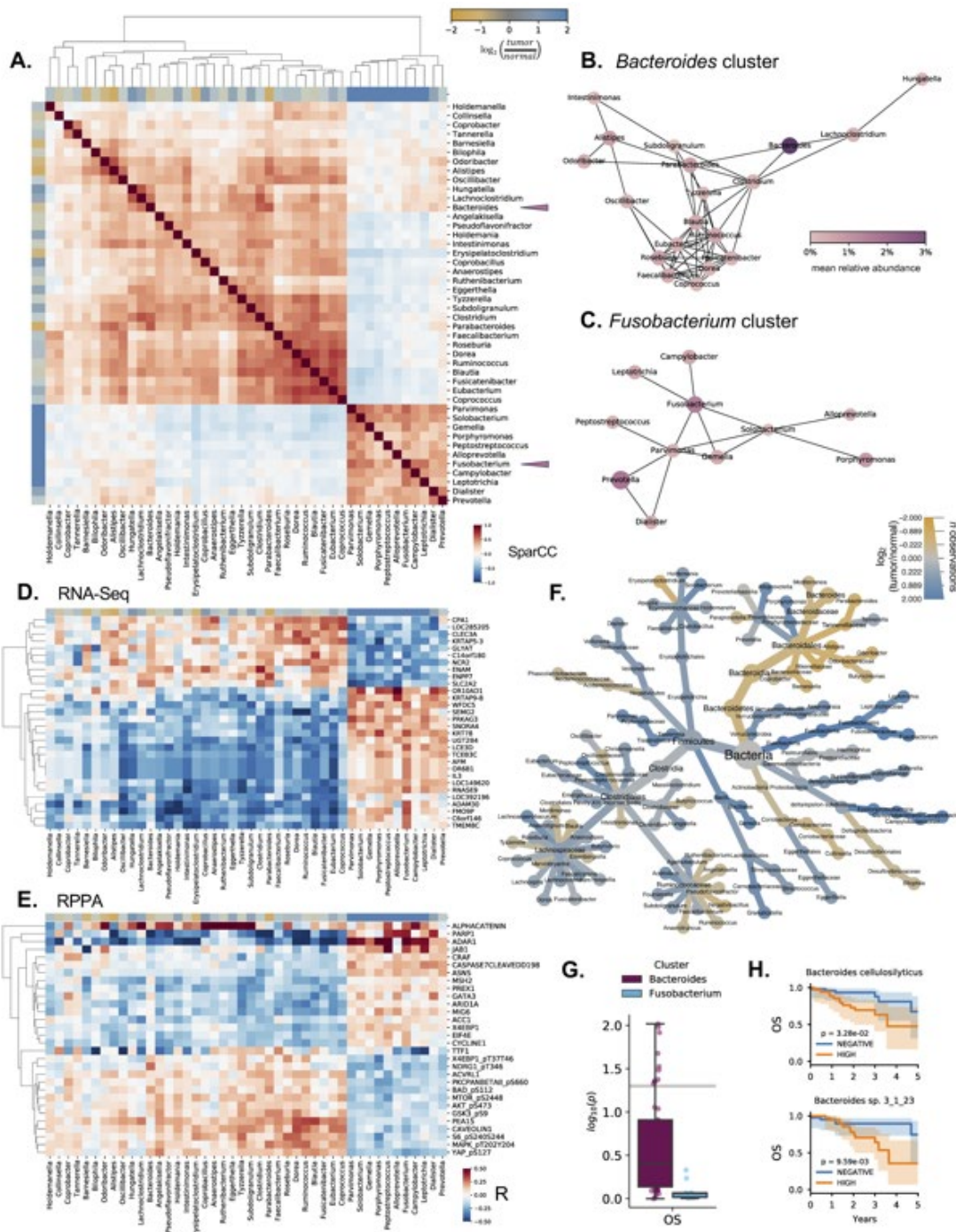


Figure 3.5: Colorectal tissue microbiomes cluster into Fusobacterium and Bacteroides coabundance groups predictive of host tissue molecular environment

- (A) Heatmap clustering of correlations between bacterial genera reveals anti-correlated clusters of genera, characterized by *Bacteroides* and *Fusobacterium* (purple triangles). Axes are colored according to species' association with tumor (blue) or matched adjacent normal tissue (yellow).
- (B – C) *Bacteroides*- (B) and *Fusobacterium*-associated coabundance networks. Nodes size proportional to the prevalence of the genera in tissue samples, and node hue is proportional to abundance.
- (D – E) Coabundance groups are predictive of gene expression (D; RNA-Seq) and protein expression (E; RPPA).
- (F) Heat-tree comparing bacterial taxa abundance in tumor samples (blue) or matched normal tissue (yellow).
- (G) Survival analysis p-values of species in the *Bacteroides* and *Fusobacterium* coabundance groups.
- (H) Overall survival (OS) curves for *Bacteroides* spp..

3.2.11 Colorectal tissue microbiomes cluster into *Fusobacterium* and *Bacteroides* coabundance groups

Having validated the contamination-adjusted microbial profiles, we sought to leverage the decontaminated TCMA dataset to investigate whether certain subgroups of microbiota were more likely to be found together in tissue from CRC patients. Using the bootstrapping procedure SparCC [155], we found two anti-correlated coabundance groups (Figure 3.5A-C, Figure S3.5A-B). The “*Fusobacterium* cluster” contains *Porphyromonas*, *Prevotella*, *Peptostreptococcus*, and *Campylobacter*, among other species that were associated with tumor samples. The second “*Bacteroides* cluster” is larger and

contains a highly correlated set of microbes, including *Parabacteroides*, *Clostridium*, and *Alistipes*. This group may represent a more normal/healthy microbiome, as several of these species were positively associated with normal tissue samples. Taxa in the *Fusobacterium* cluster were significantly associated with colorectal neoplasms, while taxa in the *Bacteroides* cluster were associated with *C. difficile* infection, irritable bowel syndrome, and cirrhosis ($q < 0.05$). These coabundance groups may represent two distinct “enterotypes” of CRC tissue microbiomes.

3.2.12 Bacterial coabundance groups are predictive of the host tissue molecular environment

We next evaluated if the bacterial coabundance groups we identified had discernible effects on host gene expression or regulation. Microbiota and host cells are known to engage with one another through a complex variety of molecular interactions. Host-derived nutrients and dietary macromolecules are utilized by microorganisms as a food source, while microbial byproducts including short-chain fatty acids (SCFAs) are known to modulate gene expression, cell differentiation, and inflammatory response [183, 184].

The TCGA database contains a dense cube of molecular profiling data, including matched genetic, epigenetic, transcriptional, and proteomic assays performed on thousands of samples. Thus, the ability to compare microbial profiles with matched host molecular profiles represents an unprecedented opportunity for querying host-microbe

interactions in various tissue types. As proof of principle, we used batch-normalized RPPA, mRNA-Seq, miRNA-Seq, and Methylation 27K data from TCGA to compute correlations between features in these datasets with genera in the *Fusobacterium* and *Bacteroides* clusters identified previously (Figure 3.5D-E, Figure S3.5C-D). For each of these assays, we found that these bacterial coabundance groups were predictive of host gene expression patterns. For instance, in RPPA protein expression data we found that ADAR1 and PARP1 expression appeared to distinguish these coabundance groups (Figure 3.5E). The protein ADAR1 is up-regulated by inflammatory mediators such as TNF-alpha and IFN-gamma [226] and regulates pathogen detection and autoinflammation by discriminating self from non-self RNA [227], while PARP1 regulates DNA repair and is activated by *Helicobacter pylori* in gastric cancer [228]. Independently, we found that ADAR1 expression was correlated with expression of PARP1, TNF-alpha, and IFN-gamma in TCGA RNA-seq data for CRC (Figure S3.5E). These results suggest that genes regulating inflammation and pathogen response may distinguish the *Fusobacterium* and *Bacteroides* coabundance groups in CRC. More broadly, these analyses illustrate the utility of TCMA as a unique resource for comparing microbial and multi-omic host profiles from matched tissue samples.

3.2.13 Matched tumor-normal analysis reveals known and novel species associated with colorectal neoplasms

The TCGA database contains detailed annotations on each tissue donor, including statistics on tumor stage, size, morphology, and location, as well information on patient survival, treatment history, and therapeutic response. To identify microbes predictive of pathological and prognostic characteristics of CRC tissue, we used matched normal tissue and primary tumor samples to perform a paired comparison of tissue-resident microbes (Figure 3.5F, Figure S3.5F-G). This analysis identified 37 species that were significantly enriched in either normal ($n = 14$) or tumor ($n = 23$) samples ($p < 0.05$) (Table S3.2).

The species most significantly associated with CRC tumors compared with matched normal tissue was *F. nucleatum* ($p = 1.82E-3$), which is known to promote intestinal tumorigenesis. Overall, approximately half of tumor-associated species belonged to the genus *Fusobacterium*, including *F. hwasookii*, *F. massiliense*, and a number of unclassified *Fusobacterium* spp. ($p < 0.01$). Non-*Fusobacterium* species associated with CRC tumors included *P. micra*, *S. moorei*, and *P. stomatis* ($p < 0.05$), several of which belonged to the *Fusobacterium* coabundance group (Figure S3.5A) and have previously been implicated in CRC [39, 229, 230]. Other species, including several *Campylobacter* spp. did not have extant links to the disease. Of these, *C. ureolyticus* ($\text{Log}_2\text{FC} = 1.97$; $p = 2.19e-2$) is an emerging gastrointestinal pathogen implicated in inflammatory bowel disease

and colitis [231, 232], prompting further examination. *C. ureolyticus* abundance was correlated with expression of several genes, including *CAMK2D* and *UGDH* (Figure S3.5JH-I), and several genes expressed by *C. ureolyticus* were significantly associated with tumor samples compared with normal tissue (Figure S3.5J). Additionally, *C. ureolyticus* was associated with worse progression-free interval (PFI) in recurrent CRC patients (Figure S3.5K).

Taxa that were significantly more abundant in adjacent normal tissue compared to matched tumor tissue were dominated by *Bacteroides* and *Parabacteroides spp.* ($p < 0.05$) (Figures S3.5G), many of which belonged to the *Bacteroides* coabundance group (Figure S3.5A). By leveraging patient-matched tumor and normal tissue samples, TCMA may thus be used identify both known and novel bacterial associations with CRC and other gastrointestinal cancers.

3.2.14 Survival analysis reveals candidate microbial biomarkers predictive of clinical outcomes

Using survival data collected by the PanCanAtlas [233], we next examined if coabundance groups were predictive of overall survival (OS). For each species, we used a log-rank test to assess its individual prognostic value. Interestingly, species in the *Bacteroides* coabundance group were generally more prognostic of survival than the *Fusobacterium* coabundance group (Figure 3.5G). We found over a dozen *Bacteroides spp.* that were prognostic of survival, including *B. cellulosilyticus* and several unclassified

Bacteroides spp. (Figure 3.5H, Figure S3.5L). These findings demonstrate the utility of TCMA for the identification of prognostic microbial biomarkers relevant to CRC and other cancers.

3.2.15 Microbial presence in CRC tissue is predictive of host immunogenic response, inflammatory cancer pathways, and cell-cell adhesion

We next explored whether the 37 species we identified as significantly associated with either tumor or normal tissue samples had identifiable effects on host gene expression or related biological pathways. Comparing normalized abundances of these species with matched mRNA expression data from 159 CRC tumor samples, we computed correlations and found transcriptional patterns that were associated with both tumor- and normal tissue-associated species (Figure 3.6A, Figure S3.6A). Given the observed differences in the transcriptional correlations of tumor- and normal tissue-associated bacteria, we next performed gene set enrichment analysis [234] to identify biological pathways associated with the abundance of these species.

Pathway analysis revealed that (1) genes correlated with the abundance of bacterial species were consistently enriched for the activation of immune system pathways and processes, irrespective of their association with tumor or normal tissue (Figure 3.6B, Figure S3.6B) and (2) processes related to inflammatory cancer pathways and cell-cell adhesion were enriched among genes correlated with tumor-associated and

normal tissue-associated species, respectively (Figure 3.6C-D, Figure S3.6C-D).

Specifically, both tumor- and normal tissue-associated species were enriched for processes relating to intestinal IgA production, antigen presentation, natural killer cell-mediated cytotoxicity, cytokine signaling, and primary immunodeficiency, suggesting near-universal activation of an immunogenic transcriptional response to the presence of these bacteria (Figure 3.6B, Figure S3.6B).

We also found that pathways including DNA replication, DNA repair, oxidative phosphorylation, p53 signaling, and ribosome activity were all negatively enriched among normal tissue-associated species, and positively enriched among tumor-associated species, particularly for *Fusobacterium spp.* (Figure 3.6C, Figure S3.6C). Conversely, genes involved in the regulation of cellular adhesion were positively enriched among normal tissue-associated species and negatively enriched among tumor-associated species (Figure 3.6D, Figure S3.6D). Together, these results indicate that within this cohort of CRC tumor samples, tumor-associated species may be associated with proinflammatory, neoplastic transformations and loss of epithelial integrity.

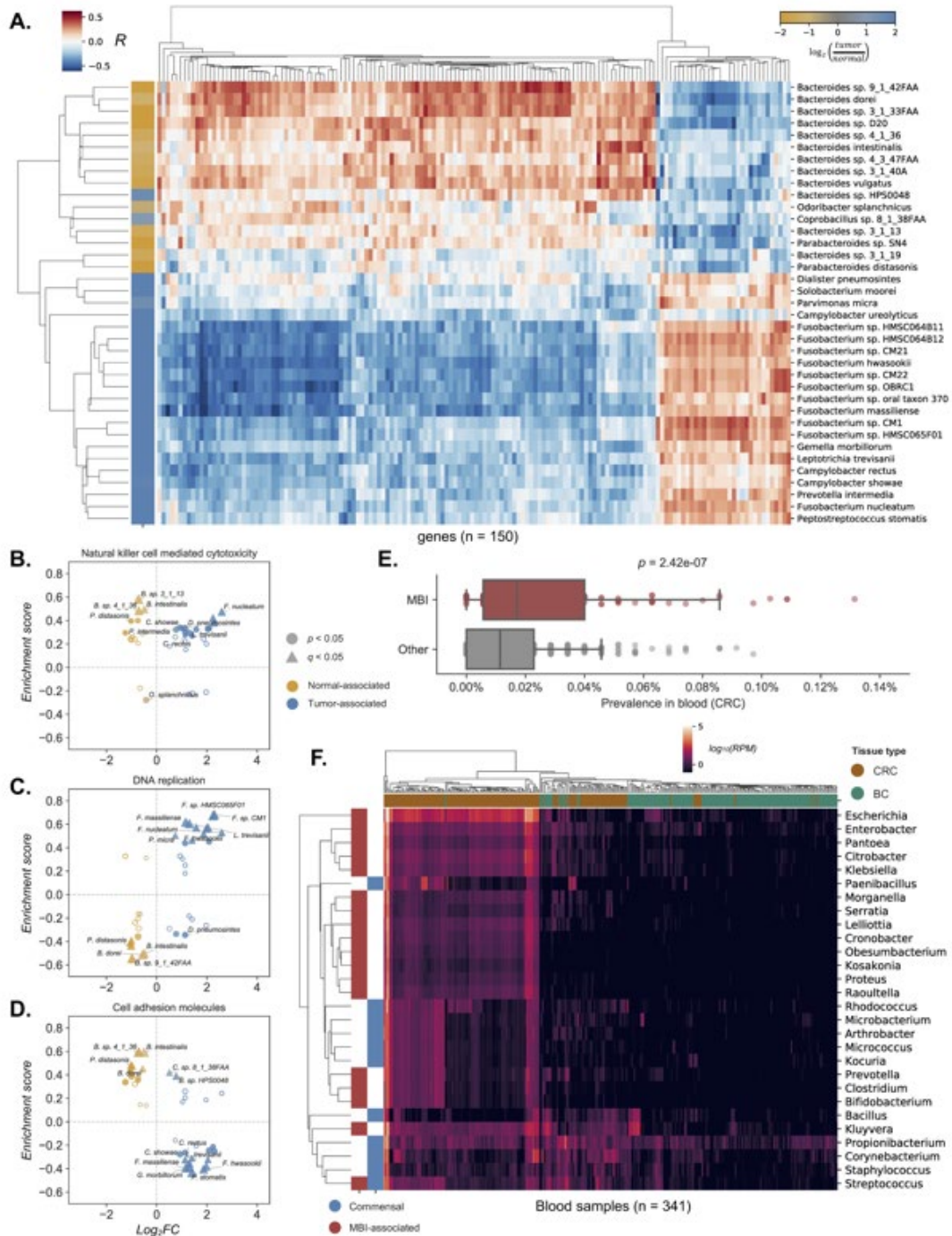


Figure 3.6: Microbial presence in CRC tissue is predictive of host gene expression pathways and mucosal barrier injury

- (A) Correlation between host gene expression (columns) and CLR-transformed species abundances (rows). Rows are colored according to each species' association with tumor (blue) or normal tissue (yellow).
- (B – D) Comparison of differentially abundant species and their association with tissue type (x-axis) versus enrichment score (y-axis) for KEGG terms “Natural killer cell mediated cytotoxicity” (A), “DNA replication” (B), and “Cell adhesion molecules” (C).
- (E) Bacterial species implicated in MBI are more prevalent in decontaminated blood samples than other species.
- (F) Bacterial genera implicated in MBI (red) are more abundant in CRC blood (brown) than BC blood (green), in contrast to some commensal species (blue).

3.2.16 Microbial presence in CRC blood samples indicate mucosal barrier injury

As shown in Figure 3.1B and Figure S3.1B, bacteria were significantly more abundant and diverse in blood samples from CRC patients than from BC patients ($p < 0.01$). The presence of transient, endogenous microbial DNA in the bloodstream has been reported in primary CRC patients, often before diagnosis, and may even be predictive of tumor stage and location [62, 235]. Loss of mucosal barrier function is a common feature of CRC and other chronic inflammatory conditions, and may lead to microbial translocations from CRC tumors to the lamina propria and bloodstream [236, 237].

To explore this possibility, we examined the abundance of subsets of bacterial species and genera designated as common commensal ($n = 407$) or MBI-associated ($n =$

693) [213]. Examining MBI-associated species among decontaminated blood samples within the CRC cohort, we found that species associated with MBI were considerably more prevalent than those that were not ($p = 2.42E-7$; Figure 3.6E). We then compared the abundance of MBI-associated genera with that of common commensals and discovered that genera associated with MBI were frequently more abundant in the blood of CRC patients than BC patients (Figure 3.6F, Figure S3.6E). These results point towards the potential utility of bloodborne bacterial DNA from MBI-associated organisms as a potential biomarker for CRC.

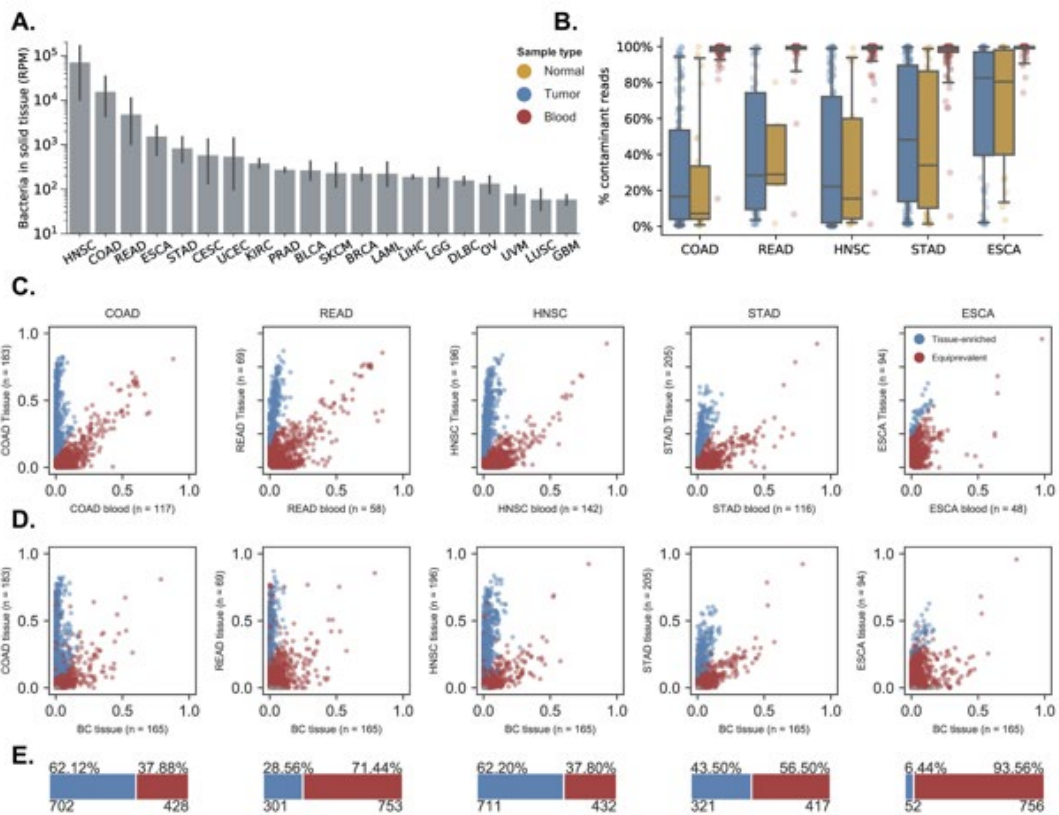


Figure 3.7: Contamination-adjusted tissue microbiome profiles for all gastrointestinal cancers in TCGA

- (A) Pan-cancer abundance of bacteria in solid tissue samples from TCGA projects prior to decontamination. Data are represented as mean \pm 95% CI.
- (B) Estimated fraction of contaminant reads for sequencing experiments on tumor (blue), normal (yellow), and blood (red) samples for each sequencing project in TCMA.
- (C) Classification of tissue-resident (blue) and contaminant (red) species across TCGA gastrointestinal tissues by comparison of prevalence in blood and tissue.
- (D) Labeling of tissue-resident (blue) and contaminant (red) species across gastrointestinal tissues by comparison of prevalence in brain tissue and disease-specific tissue, using classification from (C).

- (E) **Estimated proportions of tissue-resident (blue) and contaminant (red) species for each TCGA project.**

3.2.17 Contamination-adjusted tissue microbiome profiles for all gastrointestinal cancers in TCGA

Having successfully identified the CRC tissue-associated microbial component in the TCGA dataset, we analyzed samples from other cancer types to search for tissue-resident microbiota. All sequencing datasets contained some number of bacterial reads, but they were most abundant in gastrointestinal cancers, as expected (Figure 3.7A). In particular, tissue samples from head & neck cancer (HNSC), colon cancer (COAD), rectal cancer (READ), esophageal cancer (ESCA), and stomach cancer (STAD) had the greatest number of bacterial reads prior to decontamination, whereas uveal melanoma (UVM), lung squamous-cell (LUSC), and glioblastoma had the fewest.

Given the abundance of microbial reads in gastrointestinal cancer types, we used our prevalence-based approach to determine if tissue-resident microbiota were present and estimate the fraction of contaminant reads in each sample type (Figure 3.7B). In addition to COAD and READ, we found a strong signature of tissue-resident species in HNSC, STAD, and ESCA projects by comparing species prevalence in tissue with blood and brain samples (Figure 3.7C-D), and we estimated the fraction of minimally detectable species that were tissue-resident or contamination (Figure 3.7E). For each of

these cancer types, we applied our decomposition model to isolate tissue-resident populations and establish TCMA. Few statistically significant tissue-resident populations in bladder (BLCA), breast (BRCA), uterine (UCEC), cervical (CESC), or prostate (PRAD) cancers could be detected (Figure S3.7A-B). The microbial biomass in these tissues is known to be magnitudes less than that of gastrointestinal tissues despite similar levels of contamination ($p = 0.56$), hence it may be more challenging to distinguish the few tissue-resident species from the overwhelming proportion of contaminants.

3.3 Discussion

By comparing and integrating data from multiple NGS platforms and various sample types, we isolated and experimentally validated the tissue-resident component of these datasets, thus producing the first public resource of computationally decontaminated microbial profiles in TCGA tissue samples. This examination of equiprevalence provides a blueprint for future analyses of sequencing data for metagenomic profiling of tissue-resident microbiota. Putative contaminant species are more likely to originate from a single source and are also expected to demonstrate a lesser degree of intraspecies genetic variation, meaning that additional analyses of gene- and nucleotide-level prevalence may be helpful for controlling contamination, as we demonstrated for mixed-evidence cases such as *E. coli*. Prevalence-based analyses are

likely to supplement standard batch-correction tools, which control technical variation but do not explicitly model contamination. More statistically rigorous tools that leverage prevalence and other technical variables to explicitly define observed metagenomic data as some linear combination of endogenous and contaminant read counts, may therefore be warranted.

The ability to retroactively remove contaminant species from NGS sequencing datasets will greatly expand the breadth and accessibility of metagenomic profiles for downstream analyses. Multi-institutional initiatives such as TCGA and GTEx have collected tens of thousands of tissue samples for sequencing, many of which are from internal organs and tissue types known to harbor microbiota. Most of these samples have been characterized extensively along genetic, epigenetic, transcriptional, and proteomic axes or provide detailed clinical profiles on patient donors. Meanwhile, a growing body of evidence suggests that alterations to the microbiome are associated with cancer development, progression, and drug response [32-34]. Therefore, obtaining robust profiles of the microbial composition of human tissues in these sequencing databases will provide new insights into multi-omic host-microbe interactions in human tissue samples that would otherwise be difficult to acquire and analyze.

As proof of principle, we used TCMA to identify two dominant clusters of tissue-resident bacteria in CRC samples, as well as their associated molecular expression

patterns and prognostic significance. Pathway analysis of matched transcriptional data demonstrated that tumor-associated species were positively correlated with cancer-related inflammatory pathways and negatively associated with cellular adhesion machinery. Specifically, enrichment of ribosome, p53 signaling, DNA repair, oxidative phosphorylation, and cellular adhesion pathways may point to a previously described mechanism wherein inflammatory cytokines downregulate p53 by stimulating ribosome biogenesis in colonic epithelial cells, leading to downregulation of E-cadherin and epithelial-mesenchymal transition [238]. Given the established ability of many of these species to induce inflammation [39], stimulate cytokine activity [239], and modulate E-cadherin [240] in colonocytes, the contribution of these species to this inflammatory cancer pathway necessitates further exploration.

Beyond CRC, the TCMA database will allow interrogations of pan-cancer relationships between the microbiome and tumor development. In most cases, the role of microbiota in cancer is context-specific. For example, *H. pylori* is known to advance gastric cancers but seemingly offer a protective effect in esophageal adenocarcinoma [241]. However, certain pathogenic processes, such as chronic inflammation, altered metabolic states, and abrogation of viral latency display commonality across cancers [242]. Since TCGA samples were collected and analyzed with common methodologies, the decontaminated metagenomic profiles for thousands of tissue samples presented

here provide an ideal platform for examining host-microbe relationships that span cancer types, in contrast to meta-analyses which must integrate data from disparate sources. Thus, in addition to providing methodology for comprehensively identifying and removing contamination, TCMA represents an unprecedented resource for exploring the role of tissue-resident microbiota in various cancer types and identifying novel, predictive microbial biomarkers.

3.4 Methods

3.5 Acquisition and metagenomic profiling of TCGA sequencing data

The raw TCGA bam files and the analyte, sample, and patient metadata associated with each sequencing run were obtained from the NCI Genomic Data Commons (GDC) via the GDC's application programming interface (API). Specifically, WXS data were accessed from the GDC data repository and WGS data were accessed from the GDC's legacy archive. Overall, we acquired bam files from 19,409 sequencing runs (WGS: $n = 4,608$; WXS: $n = 15,066$) for all TCGA cancer types with WGS or WXS data available.

All WGS and WXS data from TCGA samples were screened for microbial content using the PathSeq pipeline [71], which is made available as part of the Broad Institute's Genome Analysis Toolkit (GATK 4.0). The PathSeq analysis was performed using prebuilt human and microbial reference genomes and the NCBI taxonomy database

from the PathSeq resource bundle, which were accessed via ftp from the Broad Institute in December 2017. PathSeq was used with default settings, with the exception of the minimum clipped read length, which was set to 50 to minimize the false positive rate. All sequencing data were analyzed on a local high-performance computing (HPC) cluster, which is comprised of 60 compute nodes, 1,512 CPU cores, and approximately 15TB of RAM.

Unambiguously aligned sequencing reads for bacteria at each taxonomic level were aggregated for available WGS and WXS data from 22 TCGA sequencing projects representing a total of 19,409 sequencing runs (4,608 WGS and 15,066 WXS). Total read counts for TCGA input bam files and PathSeq output bam files were calculated using SAMtools' flagstats function for RPM normalization. Total bacterial abundance values were then normalized to the total read count (in millions) of the input bam files. Aggregated PathSeq results and associated metadata for each sequencing run were then deposited as phyloseq objects for downstream analyses in R.

3.5.1 Decomposition of observed TCGA microbial profiles into tissue-resident and contaminant fractions

The classification of tissue-resident microbiota for each TCGA project was performed at the species-level using WGS sequencing data. To assess whether a species deviated significantly from equiprevalence and identify a tissue-resident population, we found the most generalizable criteria combined a statistical test of proportions with a

hard cutoff on blood prevalence. Species were defined as tissue-resident if they were prevalent in fewer than 20% of blood samples and significantly more prevalent in tissue than blood by a one-sided Fisher exact test ($q < 0.05$).

Fisher's test offered two major benefits: (1) it performs well for low prevalence cases, meaning that it naturally removed low-prevalence species which could not be statistically distinguished from contamination, and (2) it is sensitive to the group sample sizes provided (tissue and blood), making it sufficiently generalizable across sequencing projects which had varying numbers of tissue and blood samples. Because of the very large total number of detectible species ($n = 11,745$ in CRC), we used FDR-correction to adjust for multiple tests ($q < 0.05$). The second filter was hard cutoff on blood prevalence (<20%). This was effective for classifying high-prevalence species, which were statistically more prevalent in tissue than blood but still detectible in more blood samples than was plausible for endogenous blood-borne bacteria. Ultimately, for CRC data the second cutoff was only relevant for four species from the *Enterobacteriaceae* family, which likely represent mixed-evidence species. Finally, we defined "detectible" as having more than one sequencing read aligning to a given taxon (≥ 2 reads). Singletons (taxa with a single read) are known to frequently be sequencing artifacts or false positives and are commonly removed to reduce noise in downstream metagenomic analyses.

Many reads are not aligned at the species level. For example, unambiguous genus-level alignments are not necessarily equal to the sum of unambiguous species-level alignments from species within that genus. Therefore, in order to preserve read counts at taxonomic ranks above species-level, we adjusted read counts to reflect that a given clade could be comprised of a combination of contaminant and tissue-resident species. The decomposition of observed metagenomic data (K) into tissue-resident (T) and contaminant (C) components for a given taxon in a given sample can be described using a mixture with two components of the form $K = mT + nC$, where m and n represent the estimated fractions of tissue-resident or contaminant sequencing reads belonging to a given taxon, respectively (such that $m + n = 1$). For all taxa above the species-level, we assigned m and n using the relative fractions of unambiguously aligned sequencing reads from species classified as tissue-resident or contaminant within the corresponding clade. For taxa with fewer than 5 unambiguously assigned reads, we imputed mixtures from other sequencing runs processed on the same plate or center. Defining these mixtures thus allowed us to propagate the classification of tissue-resident species to higher taxonomic ranks on a sample-by-sample basis while preserving read counts that were unambiguously aligned above the species level.

3.5.2 Gene-level sequencing analysis of representative species

For each of bacterial genomes of interest, the fasta sequences and gff3 files were downloaded from NCBI. The gff3 files were converted to gtf using GffRead [243]. The fasta and gtf files were then analyzed with STAR [244] genomeGenerate to make genome files for the alignment process. For each sequencing run, the output bam file from PathSeq was used to align to annotated bacterial genomes. Using subread's featureCounts [245], the output from the STAR aligner was then used to determine the read counts for each gene. For the genome coverage analysis, outputs from STAR were sorted and converted to bam files using SAMtools [246]. Deeptools [247] bamCoverage function was then used to generate the bedgraph files using RPKM normalization. The files were intersected using bedtools [248]. The \log_{10} read counts of each of the samples were summed and divided by the total number of samples per track. These genome tracks were then plotted using the Circos software [249]. For visualization of *cadA* and *ldcC* alignments, counts from each sample and bin (10bp) were summed and divided by the total number of samples per track. The bedgraph files were converted to bigwig using UCSC bedGraphToBigWig, then bigwigs were plotted using The Integrated Genomics Viewer (IGV). Each track was scaled to the max bin height within the viewable region.

3.5.3 Nucleotide-level analysis of bacterial sequence variants

For each bacterial genome in the PathSeq reference, we screened each BAM file from the PathSeq output (COAD-READ, WGS) for sequence variants using the GATK HaplotypeCaller pipeline. Variant calling and quality filtering parameters were chosen according to previously described methodology for bacterial sequencing data [250]. The output VCF files were then converted to TSV format and were aggregated across sequencing runs. We defined each sequencing variant as a unique combination of genome accession ID, nucleotide position, reference base, and alternative base, producing a total of 3,445,630 unique variants across all genomes and sequencing datasets. For downstream analysis, this total was further filtered to select 143,215 (4%) features that were present in at least three sequencing runs. Strain-level genome accession IDs were then mapped to NCBI taxonomy IDs and associated lineage using the ete3 python package, then aggregated by species and genus for comparative prevalence analysis.

3.5.4 Acquisition and analysis of original TCGA tissue and plasma samples

For validation of TCMA we obtained original, matched tissue and plasma samples from a total of five CRC patients from Indivumed, an original TCGA tissue provider. Plasma samples from three healthy subjects were obtained from patients at Duke hospital. For tissue samples, microbial DNA was extracted from tissue samples

using the MoBio PowerMag Soil DNA isolation kit (Qiagen Cat# 27000-4-KF), following the Earth Microbiome Project (EMP) protocol (<http://www.earthmicrobiome.org/>) [251]. Microbial DNA was extracted from plasma using the QIAamp UCP Pathogen Mini Kit (Qiagen Cat# 50214) following a protocol developed by Jiang et. al [252]. Briefly, plasma samples were pre-treated with proteinase K, followed by lysing and spin-down through QIA amp UCP spin column. After washing with AW1 and AW2 buffers, microbial DNA was eluted in 50uL buffer AVE for downstream 16S library preparation and sequencing.

Bacterial compositions of isolated DNA samples were determined by amplification of the V4 variable region of the 16S rRNA gene by polymerase chain reaction using the forward primer 515 and reverse primer 806, following the EMP protocol. These primers carry unique barcodes that allow for multiplexed sequencing. Equimolar 16S rRNA PCR products from all samples were quantified and pooled prior to sequencing. Sequencing was performed on a 250bp PE MiSeq lane at the Duke University Center for Genomic and Computational Biology sequencing core. The 16S sequencing results were analyzed using QIIME2 [253]. Paired-end sequencing reads (250bp) were demultiplexed, denoised, and forward reads were trimmed at 10bp from the left and at 240bp on the right, while reverse reads were trimmed at 10bp from the left and at 220bp on the right. Taxonomic assignments were performed using the

GreenGenes database with 99% OTUs at all taxonomic levels [254]. Read counts for all observed taxa were summed over all assigned operational taxonomic units.

3.5.5 Estimation of bacterial coabundance groups and associated molecular signatures

Compositional effects in microbiome data often complicate the calculation of correlations between microbiota; we therefore used SparCC [155] to estimate taxa that are coabundant. This method relies on a bootstrapping procedure to control for spurious results common in microbiome survey data. Following the filtering criteria recommended in the SparCC paper, we removed samples with fewer than 500 reads and taxa with an average abundance of fewer than 2 reads per sample prior to calculating correlations. We ran SparCC with default parameters on decontaminated CRC tissue sequencing data for 100 iterations to identify coabundant taxa. The results of MicroPattern [255] pathway and disease-association enrichment analysis were obtained by identifying the top 20 genera most correlated with each of *Fusobacterium* and *Bacteroides*.

To estimate molecular signatures associated with these coabundance groups, we collected batch-normalized molecular profiling data from the PanCanAtlas publication page, including RPPA, miRNA-seq, mRNA-seq, and Methylation 27K experiments performed on matched TCGA samples. Prior to calculating Pearson correlations between matched samples, we performed preliminary normalizations on both molecular

profiling data and decontaminated tissue profiles. A \log_{10} transform was used to ensure RNA-seq and miRNA-seq expression profiles were normally distributed. The RPPA and Methylation 27K data were left unchanged. The relative abundances of decontaminated CRC tissue profiles were normalized using pseudocounts and a centered log-ratio (CLR) transform.

3.5.6 Identification of tumor- and normal tissue-associated microbiota

Microbes associated with tumor samples or matched normal tissue were calculated in R, using a custom paired analysis function written for metacoder [256]. We filtered decontaminated microbial compositions in TCMA by selecting taxa using filtering criteria suggested by the PhyloSeq preprocessing tutorial. Such filters are standard when preparing for downstream metagenomic analyses as they remove low-abundance and low-prevalence taxa which frequently have small means and large coefficients of variation, contributing unnecessary noise for downstream differential abundance comparisons. After adding pseudocounts, we calculated the relative abundance of microbiota for each sequencing run. Across all patients with matched tumor and normal tissue, we then calculated the median \log_2 ratio between the relative abundance of each taxa in each tissue type. Significance values were calculated using Wilcoxon's rank-sums test and corrected for false discovery rate. Taxa with significant p -values ($p < 0.05$) were selected for downstream analysis.

3.5.7 Survival analysis

We performed our survival analyses using a log-rank test, using the lifelines survival analysis python package [257]. Relative abundances of decontaminated bacterial compositions for all tissue samples belonging to a given patient were used for both models. Data on patient survival, disease-free interval, and progression-free interval were collected from the PanCanAtlas' clinical follow-up data [233]. For log-rank tests, patients were segregated into two groups: one with taxa relative abundance below the bottom quartile ("low" or "absent"), and one with above the top quartile ("high"). In many cases, particularly for species-level alignments, the bottom quartile was zero and therefore may include more than a quarter of patients. To ensure quartiles were non-equal, taxa that were present in fewer than 25% of samples were excluded from the analysis. The CPH test was performed with default parameters and 10-fold cross-validation.

3.5.8 Pathway analysis of species associated with tumors or adjacent normal tissue

We used GSEA [234] to analyze gene expression pathways associated with species of interest. For each species, we defined a continuous phenotype (cls) using CLR-transformed abundance values of decontaminated TCMA data. Using RNA-seq expression data from the PanCanAtlas as the expression dataset and gene lists obtained from MSigDB (KEGG, GO Biological Process, GO Molecular Function), we ran GSEA for

1000 iterations. Analysis was performed for 158 matched tumor samples, as well as for each subset with pathological stage (I: $n = 33$; II: $n = 60$; III: $n = 44$; IV: $n = 19$) within this cohort.

3.5.9 Quantification and statistical analysis

All statistical tests between unmatched groups were performed using a Wilcoxon rank-sums test (p -value), and all statistical tests between matched groups were performed using a Wilcoxon signed-rank test (p -value) unless otherwise specified.

Statistical tests of prevalence were performed using a one-sided Fisher's exact test.

Statistical tests of variance for microbial compositions were performed using

PERMANOVA. Statistical tests for survival analyses were performed using the log-rank

test. For multiple tests, the false discovery rate (FDR; q -value) was calculated using the

Benjamini-Hochberg method. All analyses were performed in python 3.7.1 and R 3.6.1.

P-values are indicated as follows: *, <0.05 ; **, <0.01 ; ***, <0.001 .

4 A Pan-Cancer Mycobiome Analysis Reveals Fungal Involvement in Gastrointestinal and Lung Tumors²

4.1 Introduction

Cancer is among the leading causes of death worldwide [30, 32-34, 53, 258-269]. Host-bacterial immune interactions profoundly influence tumorigenesis, cancer progression, and response to therapies [30, 32-34, 53, 258-269]. Nevertheless, the role of fungi (mycobiota) in these processes remains largely unexplored, missing a potential avenue for developing novel diagnostic and preventative strategies. Mycobiota and bacteria co-colonize the mammalian gastrointestinal (GI) tract [115, 270-278], skin epithelium [270, 279], respiratory tract [280], and reproductive organs [281], forming a complex ecosystem of microbe-microbe and host-microbe interactions with significant implications for human health. Despite comprising around just 0.1% of the microbial DNA present in the gut [200], fungal infections are responsible for more than 1.5 million global deaths per year [282, 283] and species from this kingdom have a disproportionate influence on the overall microbiome and host immunity [284].

A growing body of evidence links the human microbiome to cancer and cancer outcomes, including viruses, bacteria, and fungi [285, 286]. Although multiple studies

² This chapter is exactly reproduced from a research article of the same name authored by A.B. Dohlman, J. Klug, M. Mesko, I.H. Gao, S. Lipkin, X. Shen, and I.D. Iliev, currently under review in Cell.

have linked viruses with tumorigenesis leading to major success in preventive strategy development through mass vaccination against papilloma virus, in recent years several bacterial species have also been linked to cancer. *Helicobacter pylori* is responsible for approximately 75% of attributable risk for gastric cancer [41], and its mechanisms of promoting tumorigenesis remain under active investigation. In the lower GI tract, genotoxic *Escherichia coli* [29, 220], *Bacteroides fragilis* [287], *Streptococcus bovis/gallolyticus* [235], and *Fusobacterium nucleatum* [39, 44] have each been implicated in the pathogenesis of colorectal cancer. Common among these cancer-associated bacteria is their ability to modulate host immunity and provoke chronic inflammation, features which are proposed to contribute to their tumorigenic capacity [39, 288, 289]. However, few conclusive associations have so far linked the fungal microbiome and inflammation to cancer.

The mycobiome plays a key role in activation of innate, Type 17 and B-cell mediated immunity in the gut during health and disease [272, 290]. Recently, specific mycobiota and trans-kingdom features have been linked with colorectal cancers across cohorts [291, 292]. Epidemiological studies have long ago linked fungi and fungal toxins to GI cancer development, while uptake of bioactivated amines from fungi-contaminated food are proposed to contribute to carcinogenesis and high incidence of esophageal cancer [293, 294]. Experimental studies in mice have shown that CARD9-

deficient mice are more susceptible to colon cancer upon *Candida* colonization [295, 296]. Recently, studies have linked specific fungal species such as *Candida albicans* and *Malassezia globosa* to reduced efficacy of radiotherapy or increased tumor growth in mouse models of breast and pancreatic cancer respectively [266, 297].

Intratumoral bacteria have been associated with cancer progression [30, 32-34, 53, 258-269], while specific tumor-resident bacteria have been associated with decreased overall survival [207, 298]. Additionally, recent reports have suggested that circulating bacterial DNA is present in the blood of patients with GI cancer and may serve as a non-invasive biomarker for identifying cancer [62, 298]. Previously, we demonstrated that next-generation sequencing (NGS) data from The Cancer Genome Atlas (TCGA) contained high rates of microbial sequencing reads [62, 298] which can be leveraged to characterize the intratumoral metagenome. However, the fungal component of TCGA sequencing data has been thus far unexplored.

Analyzing multiple cancer types from TCGA, we were able to extract community profiles of tumor-associated fungi with genus- and species-level resolution, at a biologically plausible rate of roughly 1:10,000 human cells. Upon analyzing the vertical and horizontal distribution of reads aligning to these fungal genomes, mycobiome profiles were thoroughly screened to identify and remove contamination and false-positive signals via a robust computational strategy combining batch-related and

genomic decontamination models. After removing contaminant taxa, we found that fungal compositions varied by cancer type, with GI sites and non-GI sites each harboring tissue-specific mycobiota. Across GI samples, we show that *Candida albicans* and several other *Candida* species, *Saccharomyces cerevisiae*, and *Cyberlindnera jadinii* are highly abundant and associated with distinct GI tumor mycobiome communities, while *Blastomyces* and *Malassezia* species are abundant in lung and breast tumors respectively. Moreover, we demonstrate that rates of *Candida* and *Saccharomyces* are predictive of host tumor gene expression and disease state. We found that *Candida* was significantly enriched in upper GI tumor samples associated with the expression of immune & pro-inflammatory genes, including IL1B, IL6, IL8, IL17C, CXCL1, and CXCL2. In the lower GI tract, we found that *Candida* was significantly enriched in metastatic colon cancers, concomitant with deregulation of genes involved in maintaining focal cellular adhesions and epithelial barrier function. Analysis of TCGA transcriptomics data further suggested that tumor-associated *Candida* species were transcriptionally active. A culture-dependent analysis detected several living *Candida* species in an independent cohort. Additionally, the presence of *Candida* was confirmed by external ITS sequencing performed on a subset of original TCGA tumor samples we obtained. Finally, we found that *Candida* is more abundant in tumor samples versus adjacent normal tissue and that the presence of *Candida* at the tumor site predicts a significant decrease in survival rates.

Taken together, these results not only implicate *Candida spp.* in the pathogenesis of GI cancers, but also point to its potential as a therapeutic target and its relevance as a prognostic tool for predicting disease outcomes.

Finally, we provide the normalized, decontaminated mycobiota profiles we uncovered from TCGA sequencing data to the research community. This curated dataset consists of fungal community profiles from 883 sequencing runs on 767 primary tumor samples from a total of 671 individuals. As these fungal profiles are derived from TCGA tumor samples, they are accompanied by detailed histological and clinical annotations, including tumor stage and patient survival. Moreover, the TCGA mycobiomes we provide correspond to tumor samples that have been analyzed extensively using parallel genomic, transcriptomic, and epigenetic molecular profiling, meaning that these datasets can be leveraged for examining multiomic host-microbe molecular interactions. We hope that this dataset will serve as a useful resource for future studies of the human mycobiome in cancer.

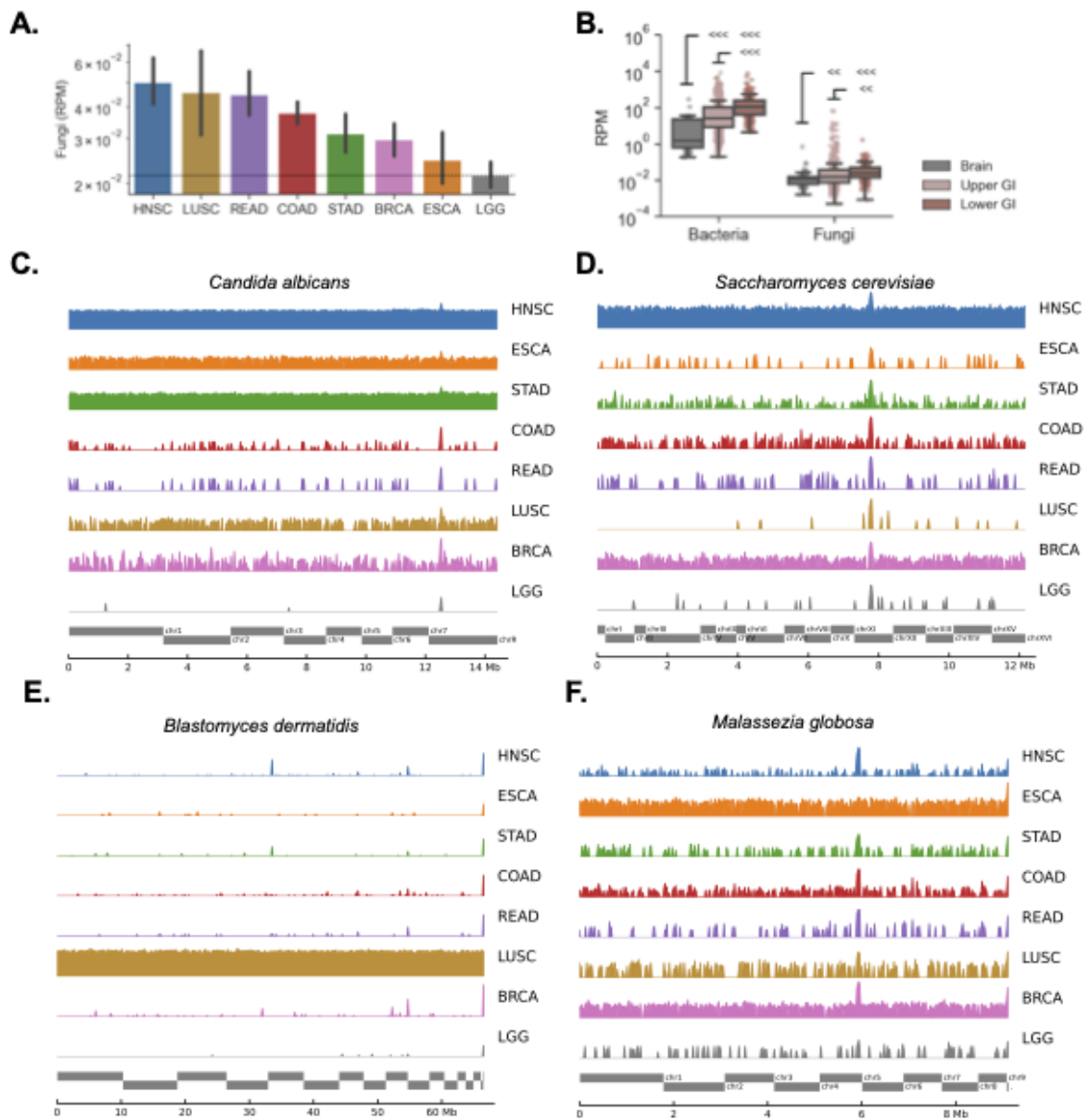


Figure 4.1: Fungal DNA is present in multiple cancer types not explained by contamination

- (A) Geometric mean of reads per million (RPM) of fungal DNA detected in tumor and tumor-associated tissue samples from head & neck (HNSC), lung (LUSC), rectum (READ), colon (COAD), stomach (STAD), breast (BRCA), esophageal (ESCA) and brain (LGG) cancers.

- (B) Both bacteria and fungi were more abundant in the lower GI tract (COAD, READ) than the upper GI tract (HNSC, ESCA, STAD), and were more abundant in both GI groups compared to brain (LGG).*
- (C) The distribution of sequencing reads aligning to the *M. globosa* genome displays similar depth across sequencing projects including brain, but reads are distributed randomly, a signature of biological contamination.
- (D) The distribution of sequencing reads aligning to the *A. bisporus* genome displays uneven depth, but reads are horizontally distributed in a predictable manner, a signature of false-positive assignments.
- (E) Genome alignments to *C. albicans* are present at high rates across sequencing projects, particularly in upper GI tumors, yet are absent in brain.
- (F) Genome alignments to *B. dermatidis* are found at high rates in lung tumors, but not elsewhere.

4.2 Results

4.2.1 Fungal DNA is abundant in gastrointestinal tumor samples from TCGA

To explore tumor-associated mycobiomes across different cancers we employed a metagenomic analysis of whole-genome sequencing (WGS) data from multiple tumor samples across different cancers available in TCGA. We selected cancer types based on previously reported presence or absence of mycobiota [30, 32-34, 53, 258-269], including gastrointestinal (GI) tissues (head-neck/HNSC, $n = 338$; esophagus/ESCA, $n = 142$; stomach/STAD, $n = 321$; colon/COAD, $n = 300$; rectum/READ, $n = 127$), non-GI external sites (breast/BRCA, $n = 229$), as well as non-GI internal sites (lung/LUSC, $n = 100$;

brain/LGG, $n = 183$), and used PathSeq [71, 299] to determine their fungal composition. The mycobiomes we detected in these tissues were then screened and filtered for contamination (See “Identification and removal of contaminant fungi and false-positive signals”).

This approach led to the detection of fungal sequences across multiple cancer patient’s tissue types, with higher rates of fungal DNA in tissues of the lung and specific sites of the gastrointestinal (GI) tract (Figure 4.1A). Across the GI tract, fungal DNA was particularly abundant in tissues from head and neck (HNSC), colorectal (COAD and READ) and stomach (STAD) tissues, and less abundant in the esophagus (ESCA) (Figure 4.1A, Figure S4.1A). By contrast, few fungal sequences were detected in brain tissue (LGG), consistent with its anatomical positioning away from barrier surfaces where fungi most frequently reside. As the brain is canonically classified as a sterile organ (fungal brain infections are usually lethal), we reasoned that bacterial and fungal sequencing reads detected in sequencing data from brain tissue likely represented biological contamination and/or false-positive signals, suggesting that such tissue can be used as a presumptive “negative control” for identifying spurious signals in other datasets [298]. Overall, samples from lower GI tissues harbored a greater density of fungi than upper GI tissues did, in a pattern consistent with bacteria (Figure 4.1B). As expected, fungal sequences represented a much smaller proportion of microbial

sequences in tissues when compared to bacterial DNA (Figure 4.1B), consistent with previous reports of intestinal human samples [271, 272, 274, 275, 291, 292, 300-302].

4.2.2 Identification and removal of contaminant fungi and false-positive signals

The discovery of fungal sequences in multiple tumor types prompted us to examine their origin, as contamination is a plausible source of fungal DNA. Microbial contamination is pervasive in metagenomic profiling experiments and can come from a variety of sources, including the laboratory environment or nucleic acid extraction kits [74, 75]. Additionally, the incorrect assignment of microbial or non-microbial sequencing reads can lead to reporting of spurious signals [303]. Thus, identification and removal of non-endogenous taxa is a necessary step that must precede downstream analyses, particularly in studies of low biomass tissue sites [76]. To ensure accurate capture of the endogenous mycobiome of these samples, we applied a rigorous, multi-step batch correction and quality control analysis to identify and remove contaminant fungi and false-positive signals from the dataset, leaving only fungal species for which there was substantial evidence of their involvement in the tissue.

We began by applying a prevalence-based decontamination model to identify and remove (1) fungal species and genera whose presence was associated with specific sequencing batches and could not be explained by biological variation, and (2) samples from multi-well sequencing plates with strong evidence of contamination. Briefly, we

used Chi-squared test to determine if taxa were overrepresented in certain multi-well aliquot plates but not in certain tissue types (See Methods). This analysis identified 23 species and 12 genera that met these criteria, including *Beauveria* and *Pochonia spp.*, which are not known to colonize humans (Table S4.1). Additionally, we removed 18 samples from a single sequencing plate which displayed evidence of significant fungal contamination (Figure S4.1B).

While tracking the presence of taxa across sequencing batches can effectively identify contaminants, such a strategy is unable to identify contamination events that span sequencing batches, nor is it capable of identifying signals which may be the result of false-positive alignments or incorrect taxonomic assignments. To address these possibilities, we performed a genome-wide analysis of sequence alignments for each of the fungal species detected in each tumor type (See Methods, Table S4.1). For each sequencing project, we compared the genome coverage depth (“Vertical QC model”) as well as the distribution of sequencing reads across the length of each genome (“Horizontal QC model”). The use of orthogonal models in this case allows for the identification of different categories of false-positive signals. Species that are truly present at the time of sequencing, but not in the original biopsies are referred to as biological contaminants and are likely to have similar levels of coverage depth across tissue types, yet a random distribution of read alignments across the span of their

genome. Conversely, false-positive alignments are likely to occur in conserved genes or highly mobile genes belonging to different fungal or non-fungal DNAs [303, 304] generating similar patterns of sequence alignments across tissue types.

For example, these analyses found that reads aligning to *Malassezia restricta* and other *Malassezia spp.* genomes displayed similar coverage depth across sequencing projects but a horizontal read distribution that was generally random (Figure 4.1F, Figure S4.1C, Table S4.1), a signature consistent with biological contamination. *Malassezia spp.* are frequently found on the skin surface [305, 306] and were likely transferred to multiple samples during the handling of TCGA biospecimens. Consistently, several *Malassezia* species were selected by this model as a true signal in skin-adjacent breast tumors (Table S4.1): a finding consistent with known *Malassezia* colonization at skin sites [305, 306]. Meanwhile, reads aligning to the genome of *Agaricus bisporus* (common mushroom or portabello) displayed a consistent horizontal distribution pattern across sequencing projects (Figure S4.1D, Table S4.1). Thus, *Malassezia restricta* and *Agaricus bisporus* were respectively removed by our vertical and horizontal QC models (Table S4.1).

Overall, our decontamination and quality control analyses resulted in the removal of 97.27% of species detected in GI tumors, 99.26% of species detected in lung tumors, and 95.53% of species detected in breast tumors. Remaining were a set of

commensal and pathogenic fungi, including *Candida albicans*, *C. tropicalis*, *C. dubliniensis*, *C. glabrata*, *C. lusitaniae*, *C. guilliermondii*, *Cyberlindnera jadinii* (Figure 4.1C, Table S4.1) and food-associated *Saccharomyces cerevisiae* and *Pichia membranifaciens* (Figure 4.1D, Table S4.1) which were abundant in GI tumors and *Blastomyces dermatidis/gilchristii* (Figure 4.1E, Table S4.1) which was abundant in lung tumors and are known causative agents of blastomycosis, a disease that primarily affects the lungs [283, 307]. Many of the species classified as contaminants and/or false-positive signals were not known to colonize humans, including plant pathogens *Alternaria alternata*, *Bipolaris oryzae*, and *Fusarium verticilloides* (Table S4.1). Notably, *Malassezia* spp. were classified as probable contaminants in all tumor types except for breast tissue, suggesting that some of these reads may be biologically relevant, as epidermal tissue is often involved in such cancers (Figure 4.1F, Table S4.1) [308]. Thus, we surmise that the detection of reads from *Malassezia* spp. in breast tissue may have originated from both endogenous and contaminant sources, as we have previously shown for *E. coli* in CRC samples [298]. Finally, we validated the abundance of several of these species in a secondary metagenomic analysis using TaxaTarget [309], a tool specifically designed for the detection of eukaryotic marker genes (Figure S4.1E).

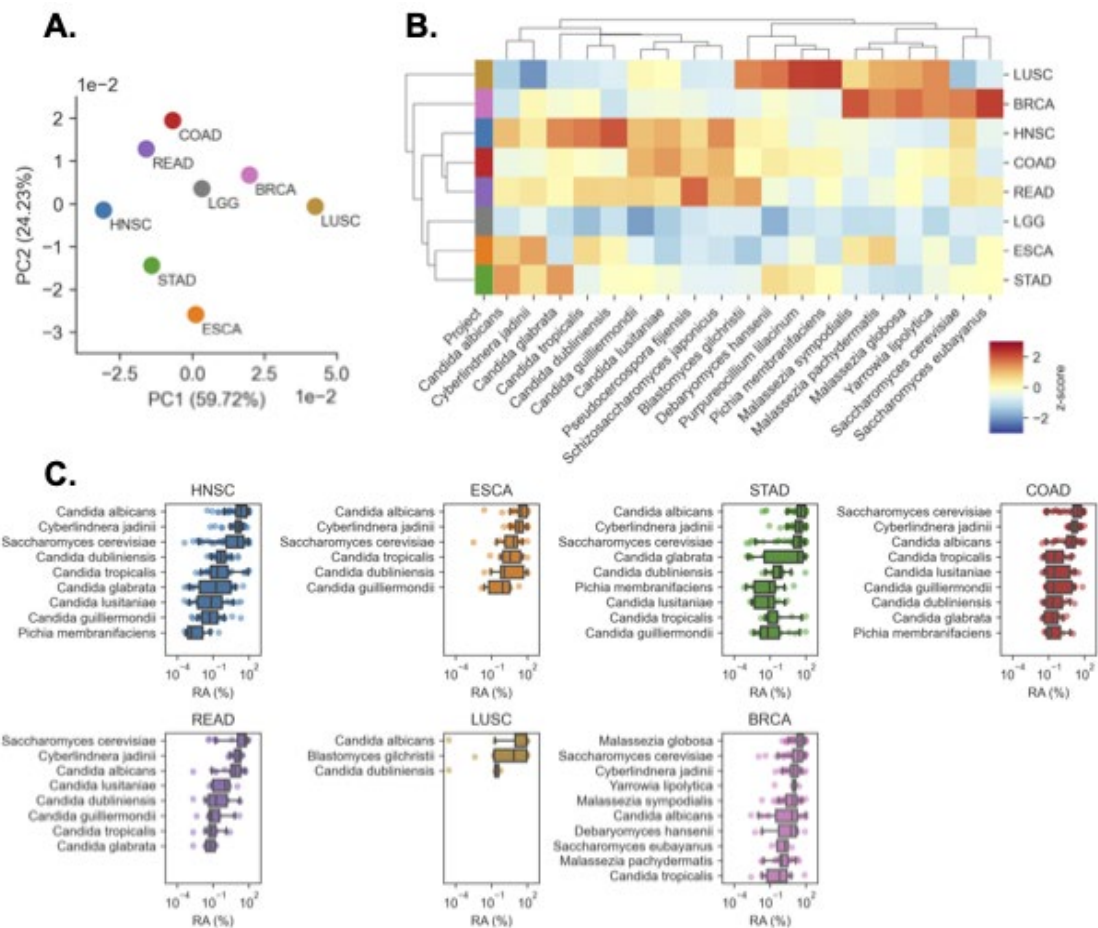


Figure 4.2: Primary tumor samples harbor disease-specific microbiomes

- (A) Principal coordinate analysis (PCoA) of normalized species abundances from head-neck (HNSC), esophageal (ESCA), stomach (STAD), colon (COAD), rectal (READ), lung (LUSC), breast (BRCA), and brain (LGG) reveal clustering by tumor type, after filtering out potential contaminants and false positive signals.
- (B) Hierarchically clustered heatmap showing difference in normalized fungal species abundances (RPM) between tissues from each TCGA cancer type, after filtering out potential contaminants and false positive signals. Species are included if they were classified as tissue-resident in any of GI, lung, or breast samples, even if they were classified as contaminants in others. Heatmap values are z-scored by species abundance.

- (C) **Boxplots showing distribution of relative abundances (RA) from the 10 most abundant species detected in each TCGA cancer type after removing potential contaminants, provided they were detected in at least 5 samples.**

4.2.3 TCGA tissue samples are composed of disease-specific fungi

Our approach generated species-level resolution data allowing the identification of specific fungi across various tumor types. Principal coordinate analysis (PCoA) and hierarchical clustering of genus abundances across TCGA cancer types revealed that head-neck, colon, and rectal tumors had highly similar fungal compositions, as did stomach and esophageal tumors, while the fungal composition of non-GI tumors were largely distinct (Figure 4.2A-B). Differences in the fungal communities we observed across GI sites could be affected by variations in pH, oxygen availability, or bacterial biogeography across the GI tract, among a few key factors driving microbial variation. The stomach is known to be a highly acidic environment (pH 1.5 - 3.5), while esophageal tissues are subject to sudden, transient drops in pH (from 7.0 to below 4.0 during reflux) [310]. By contrast, oropharyngeal, colon, and rectal tissues are characterized by a more consistent, neutral pH (6.0 - 7.0).

In addition to environmental factors, the detection of fungal species in these samples is affected by the availability of reference genomes. As such, there may be additional unknown fungal species not detected in our analysis. Nevertheless, we found

that tumor-associated fungal communities were characterized by high abundance and prevalence of *Saccharomycetales* taxa, including *Candida* and *Saccharomyces*, consistent with previous gut mycobiome studies relying on metagenomics [300], culturomics and ITS-amplicon sequencing [271, 272, 274, 275, 301, 302, 311]. In addition to these more common fungi, deeper analysis revealed the presence of sequences from multiple fungal species and genera as well as their distribution across different cancer types (Figure 4.2C, Figure S4.2A, Table S4.2).

The growing consensus on the importance of intestinal mycobiota has prompted the investigation of (1) which fungi are capable of surviving, residing, and replicating in the GI tract (fungal symbionts or commensals) to influence the host over a prolonged period, and (2) which are transient passengers, contaminants, or represent environmental fungi (non-commensal fungi) that are normally benign but can impact immunosuppressed individuals [312, 313]. *Candida spp.* were more abundant across the GI tract as compared to other body sites, consistent with their known commensal status in this part of the body [314] and ability to expand [271, 272, 274, 275, 311, 315, 316] or breach the GI barrier [276, 317] during disease (Figure 4.2C, Figure S4.2A). Species-level analysis determined that *C. albicans* was the most abundant representative of the *Candida* genus: *C. albicans* was highly abundant in multiple cancers but was particularly abundant in cancers of the GI tract (Figure 4.2B-C), consistent with previous studies

[272, 274, 275, 311, 315, 316]. *C. tropicalis*, *C. dubliniensis*, *C. glabrata*, *C. lusitaniae*, *C. guilliermondii*, *C. parapsilosis*, and *Pichia membranifaciens* were also present, but at lower abundance and prevalence across samples (Figure 4.2B-C, Figure S4.2A). *Saccharomyces* spp. were primarily represented by *S. cerevisiae* (Figure 4.2B-C) [274, 275, 300]. Among fungi broadly assigned as non-commensal, we also detected *Cyberlindnera jadinii* in multiple GI tissues (Figure 4.2B-C), a species which is found in processed food products and rarely infects people, presumably arriving via diet [318]. Lung tissues carried *B. gilchristii/dermatitidis* (Figure 4.2B-C), which are causative agents of blastomycosis [283, 307]. Interestingly, we detected evidence of *Blastomyces* DNA in 6 out of 50 patients with squamous cell lung carcinomas. In the general population, the incidence of blastomycosis is 1-2 cases per 100,000 [319]. Together, these findings indicated the presence of biologically meaningful associations linking the presence of fungal DNA to tissues from specific body sites.

- (A) Hierarchically clustered heatmap showing co-abundance among fungal species using correlation statistic SparCC reveals clusters of species *Candida*- and *Saccharomyces*-associated species across GI tumor samples. Purple boxes indicate clusters forming *Candida*- and *Saccharomyces*-associated tumor co-abundance groups.
- (B – D) Hierarchically clustered heatmaps, showing gene expression patterns in head-neck (HNSC), stomach (STAD), and colon (COAD) cancers. Heatmaps are clustered by row, while column clustering is determined by (A). Gray columns indicate species not detected in certain cancer types
- (E - G) SparCC correlation between *Candida* and *Saccharomyces* and bacterial genera found in matched tumor samples from TCMA (Dohlman et al., 2020)

4.2.4 Emergence of *Candida* and *Saccharomyces* co-abundance groups is associated with gastrointestinal cancers

Microbiota participate in a complex web of interspecies ecological interactions and the dynamics of these interaction networks can profoundly influence human health [110, 320]. To explore the potential presence of fungal interaction networks and clusters of co-abundant taxa, we applied a bootstrapping procedure SparCC [155] to analyze the mycobiota across GI cancers. This analysis uncovered that *C. albicans* and *S. cerevisiae* were each at the center of two anticorrelated co-abundance clusters which were observed across GI cancer types (Figure 4.3A). The co-abundance group associated with *C. albicans* included *C. dubliniensis*, *C. tropicalis*, and *C. guilliermondii*, while the abundance group associated with *S. cerevisiae* was comprised of taxa including *S. eubayanus*, *C. jadinii*, *P. membranifaciens*, as well as *C. parapsilosis* and *C. glabrata*. Additionally, we

found that these two co-abundance clusters were largely predictive of host gene expression across head-neck, stomach, and colon cancers (Figure 4.3B-D). These findings suggested that cancers of the GI tract may segregate into *Candida*- and *Saccharomyces*-associated tumors. Notably, many of these species in each of these clusters are taxonomically related, thus the degree to which they are driven by biological or phylogenetic factors (or both) warrants further exploration.

4.2.5 Trans-kingdom analysis reveals co-abundance groups associated with *Candida* and *Saccharomyces* in GI cancers

To further explore the microbial communities associated with the *Candida* and *Saccharomyces* tumor co-abundance clusters and their relevance to disease, we examined bacterial populations associated with *Candida* and *Saccharomyces* and applied the same correlation approach to identify associations among GI tumor-resident fungi and matched, decontaminated, intratumoral bacterial communities from The Cancer Microbiome Atlas (TCMA) [298]. This analysis identified several interesting bacterial subpopulations that were correlated with *Candida* and *Saccharomyces* in each cancer type.

In head-neck tumors, *Candida* and *Saccharomyces* were associated with similar bacteria (Figure 4.3E, Figure S4.3A). *Lactobacillus spp.* and especially *Lactobacillus gasseri* were very frequently found in the presence of *Candida* and, to a lesser extent, *Saccharomyces* (Figure S4.3D-F). This observation is consistent with reports that *Lactobacillus spp.* interact extensively with *Candida* to influence its pathogenicity [321-

323]. *Bifidobacterium*, which is known to support intestinal barrier function [324] was also positively associated with *Candida* in head-neck cancers. In contrast, we found that species associated with periodontal disease, including *Fusobacterium* spp. and *Prevotella* spp., were negatively associated with *Candida* and *Saccharomyces* in head-neck tumors.

In stomach tumors, we also observed that *Candida* was strongly associated with *Lactobacillus* (Figure 4.3F, Figure S4.3B,E). However unlike in head-neck cancer, *Candida* and *Saccharomyces* in stomach tumors were largely associated with dissimilar clusters of bacteria. Most notably, we observed that *Candida*-associated tumors were less likely to harbor *Helicobacter pylori*, which is believed to be a causative agent in many stomach cancers [41, 325, 326]. Conversely, *Saccharomyces* was more likely to be found alongside *H. pylori*. A similar pattern was identified for the genera *Streptococcus* and *Clostridium*, which were positively associated with *Candida* and negatively associated with *Saccharomyces*. Together, these results suggest that *Candida* and *Saccharomyces* may occupy similar ecological niches among bacterial communities in head-neck tumors but are associated with very different bacterial populations in stomach cancer.

In lower GI tumors, *Candida* and *Saccharomyces* were also co-abundant with distinct bacterial populations (Figure 4.3G, Figure S4.3C). Unlike upper GI cancers, we did not observe any association between *L. gasseri* and *Candida* in colon tumors (Figure S4.3F). However, we found that among colon cancers, *Candida* was positively associated

with *Dialister*, and was negatively associated with *Ruminococcus*, *Akkermansia muciphila*, and *Barnesiella intestinihominis* (Figure 4.3G, Figure S4.3C). *Ruminococcus* spp. are known to be less abundant in people with inflammatory bowel disease [327] and may play a role in degradation of starch in the colon [328]. Interestingly, *A. muciphila* is known for its anti-inflammatory properties [329] and its ability to promote healthy barrier function [330], while *B. intestinihominis* is associated with prolonged cancer survival and has been shown to modulate tumor immunosurveillance [331]. Thus, *Candida* appears to be negatively associated with several commensal species which help to promote anti-inflammatory and anti-cancer pathways in the lower intestine. *Saccharomyces* was not associated with the same bacteria as *Candida*, but was instead positively associated with *Porphyromonas*, *Leptotrichia*, and negatively correlated with *Odoribacter splanchnicus*. Interestingly, the presence of *Candida* and *Saccharomyces* were also associated with differing species of *Fusobacterium* spp. in colon cancer (Figure S4.3C), which have been shown to promote tumor development in colorectal cancer by provoking inflammation and host immune response [28, 39, 207].

Together, these findings demonstrate that *Candida* and *Saccharomyces* are associated with very diverging bacterial communities in stomach and colon tumors, but similar communities in oropharyngeal cancers. In the stomach, *Candida* and *Saccharomyces* were predictive of the presence of *H. pylori*, while in the colon *Candida* was

negatively associated with several bacteria known to promote beneficial host-microbe interactions. Additionally, we found significant co-occurrence between *Lactobacillus* and both *Candida* and *Saccharomyces* in head-neck tumors, while the association between *Lactobacillus* was comparatively weaker or absent in stomach and colon tumors (Figure S4.3D-F). In addition to providing insight into tumor-associated microbiomes, such trans-kingdom ecological interactions may be relevant for disease detection and potentially inform strategies for modulating tumor microbiomes for therapeutic benefit.

4.2.6 *Candida* and *Saccharomyces* are predictive of gene expression patterns in GI cancers

To better understand the effect of *Candida* and *Saccharomyces* co-abundance groups on GI cancers, we next sought to compare the rates of *Candida* and *Saccharomyces* across GI tumors. Due to compositional effects in metagenomic data, log-ratios between taxa represent a robust and reliable way to estimate biologically meaningful fluctuations in microbiomes [332, 333]. Across cancer types, we discovered that *Candida*-to-*Saccharomyces* ratios displayed striking bimodality, corroborating our previous observations of *Candida* and *Saccharomyces* co-abundance clusters and suggesting that GI tumors could be reliably organized into subgroups of *Candida*- and *Saccharomyces*-associated cancers (Figure 4.4A, Figure S4.4A). To understand the relevance of these two subgroups, we divided GI tumors into *Candida*-dominant (*Ca*-type) and *Saccharomyces*-dominant (*Sa*-type) clusters and compared them.

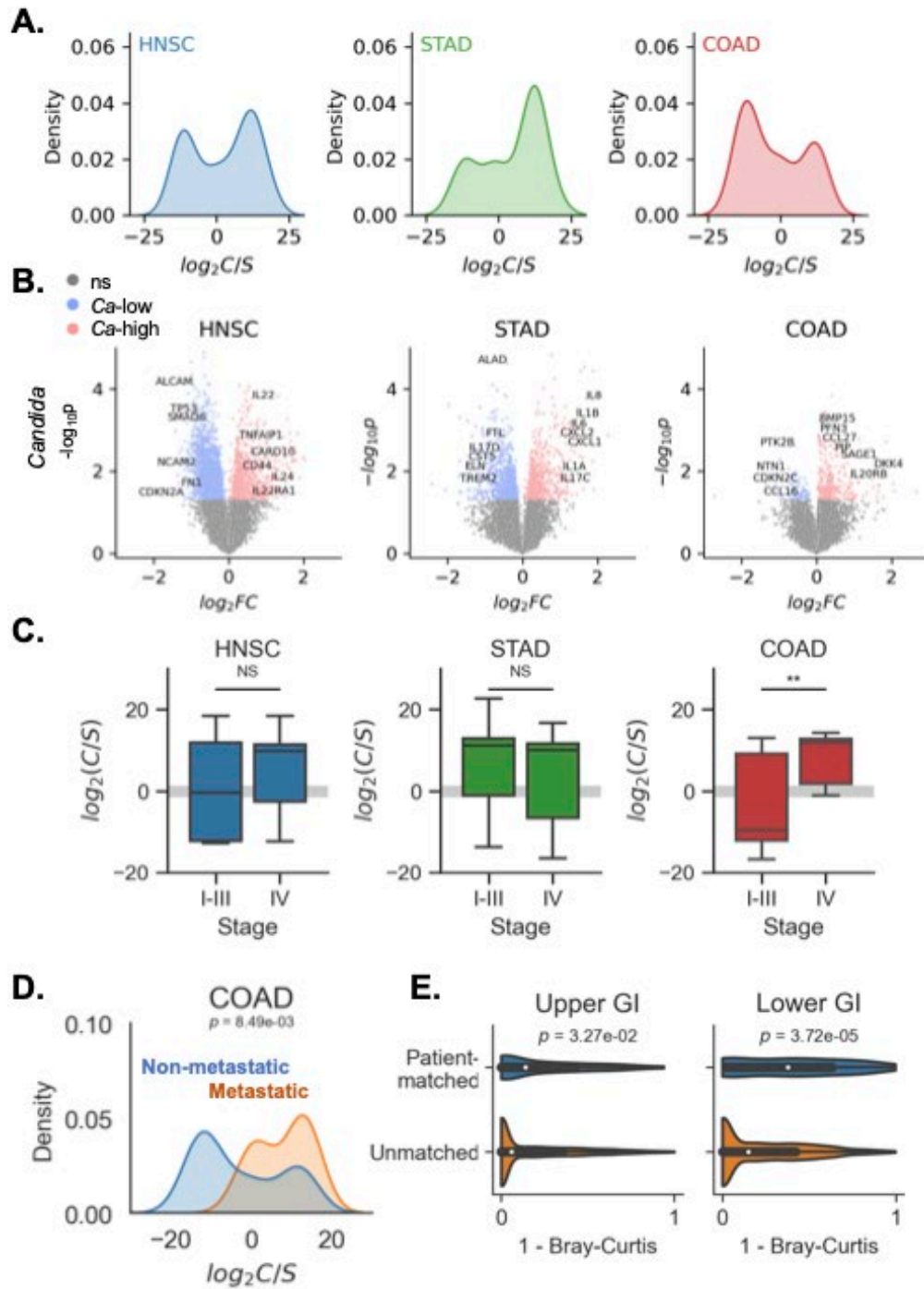


Figure 4.4: Candida is associated with late-stage and metastatic GI cancers

- (A) Kernel density estimation (KDE) of *Candida*-to-*Saccharomyces* ratios in head-neck (HNSC), stomach (STAD), and colon (COAD) cancers, suggesting that GI tumor samples can be classified into *Candida*- and *Saccharomyces*-associated cancers.
- (B) Volcano plot showing genes differentially expressed *Candida*-negative (blue) and *Candida*-high (red) tumor samples head-neck, stomach, and colon cancers.
- (C) Boxplots depicting the ratio of *Candida*-to-*Saccharomyces* in early-stage (I-III) and late-stage (IV) for head-neck, stomach, and colon cancers reveals enrichment of *Candida* in late-stage colon cancers.
- (D) KDE analysis of *Candida*-to-*Saccharomyces* ratios in metastatic (orange) and non-metastatic (blue) tumor samples finds that *Ca*-type colon tumors are significantly more likely to be metastatic.
- (E) Violin-plots showing the distribution of Bray-Curtis distances between the fungal species compositions of patient-matched tumor and blood sample (blue) and unmatched tumor and blood samples (orange). Matched samples were significantly more likely have similar composition than unmatched samples.

To see if *Ca*-type and *Sa*-type tumors harbored functional differences, we used RNA-seq data from TCGA to analyze gene expression between tumor samples that were highly abundant in *Candida* or *Saccharomyces* with tumors in which these taxa were not detected (Figure 4.4B, Figure S4.4B). This analysis identified several interesting changes in gene expression that were associated with *Candida* status. In head-neck cancer, we found that tumor-suppressors TP53 and CDKN2A were expressed at lower rates in *Ca*-type tumors, along with fibronectin (FN1) which is a marker of epithelial-to-mesenchymal transition (EMT) in head-neck cancers. Interestingly, we also saw that

IL22, IL24, CARD10, and CD44 were up-regulated in *Ca*-type tumors. These genes were not differentially expressed in *Saccharomyces*-associated tumors. Gene-set enrichment analysis (GSEA) of this expression signature demonstrated that the presence of *Candida* was associated with decreased expression of genes relating to cell adhesion molecules ($q < 0.001$) in head-neck cancers. In stomach cancers, we found that several genes related to cytokine interactions, host immunity and inflammation were positively enriched in *Ca*-type tumors, including IL1A, IL1B, IL6, IL8, CXCL1, CXCL2, and IL17C. This pro-inflammatory immune signature is consistent with previous reports that *C. albicans* invokes IL-1 β , neutrophils and Th17 cell infiltration in the gut [311]. By contrast, these genes were differentially expressed to a lesser extent or were not differentially expressed at all in *Sa*-type tumors with high rates of *Saccharomyces*. Genes down-regulated in *Ca*-type tumors included ALAD, FTL, IL17D, CST5, ELN, and TREM2. Overall, GSEA showed that this gene expression pattern was associated with significant up-regulation of genes involved in cytosolic DNA sensing ($q = 0.008$), Toll-like receptor ($q = 0.033$) signaling, Nod-like receptor ($q = 0.033$) signaling, and cytokine-cytokine receptor interactions ($q = 0.035$). In colon cancers, we found that tumor suppressor genes and genes regulating cellular adhesion pathways were downregulated in *Ca*-type tumors, including PTK2B, CDKN2C, and NET1, while genes such as BMP15, PFN3, CCL27, PIP, and SAGE1 were up-regulated in *Ca*-type tumors. Moreover, GSEA identified significant

down-regulation of genes involved in ECM-receptor interactions ($q = 0.036$) and focal adhesion ($q = 0.101$) pathways in *Ca*-type colon tumors.

These findings indicated that the presence of *Candida* in head-neck and colon tumors is associated with pro-tumorigenic and cellular adhesion-related gene pathways, while *Candida* appears to be associated with a robust immune response in stomach tumors, consistent with previous reports that *C. albicans* is linked to immune dysfunction and damage to intestinal macrophages and epithelium during pathophysiological conditions such as inflammatory bowel disease [311, 334]. However, additional analyses are needed to determine whether *Candida* plays a causative role in these gene expression changes or is merely responding to them.

4.2.7 A *Candida*-to-*Saccharomyces* ratio is associated with late-stage, metastatic colon cancer

The observation that the presence of *Candida* is associated with down-regulation of genes involved in cellular adhesion pathways and epithelial barrier function in head-neck and colon tumors led us to explore if rates between these two genera were predictive of cancer outcomes. Interestingly, *Candida*-to-*Saccharomyces* ratios were generally low among early-stage colon cancers but were dramatically increased in stage IV disease (Figure 4.4C). However, *Candida*-to-*Saccharomyces* ratios did not vary significantly by stage in head-neck, stomach, or other cancers (Figure 4.4C, Figure S4.4C). The association with late-stage colon cancer led us to examine rates of metastases

among *Ca*-type and *Sa*-type tumors. Comparing *Candida*-to-*Saccharomyces* ratios in metastatic and non-metastatic groups, we found that *Ca*-type colon tumors were significantly more likely to be metastatic than tumors with higher rates of *Saccharomyces* (Figure 4.4D; $p = 8.49E-3$; $q = 0.051$). Similar analysis did not find significant differences in other cancer types (Figure S4.4D). Thus, *Candida*-to-*Saccharomyces* ratios may capture a clinically relevant shift in tumor mycobiomes with potential prognostic value for colon cancer.

Our observation that tumor mycobiomes were predictive of metastatic colon cancer and deregulation of genes involved in epithelial barrier function led us to question if fungi or fungal DNA might transfer into the bloodstream from the barrier surfaces in which these fungi normally reside. To explore this possibility, we examined the composition of patient-matched tumor and blood samples from cancer types of the lower and upper GI tracts. We found that there were statistically significant similarities in the composition of patient-matched tumor and blood samples from patients with upper GI cancers ($p = 3.27E-2$) and lower GI cancers ($p = 3.72E-5$) compared to unmatched samples (Figure 4.4E); while the same was not true for other tumors, suggesting that the GI tract might be a possible entrance point for fungi or fungal DNA into the bloodstream. Together these data indicate that *Candida* may be associated in loss of epithelial barrier function, metastasis, and the translocation of fungi from the GI tract

into the bloodstream. However, whether *Candida* or other fungal DNA can consistently be detected in the blood of GI cancer patients requires additional examination.

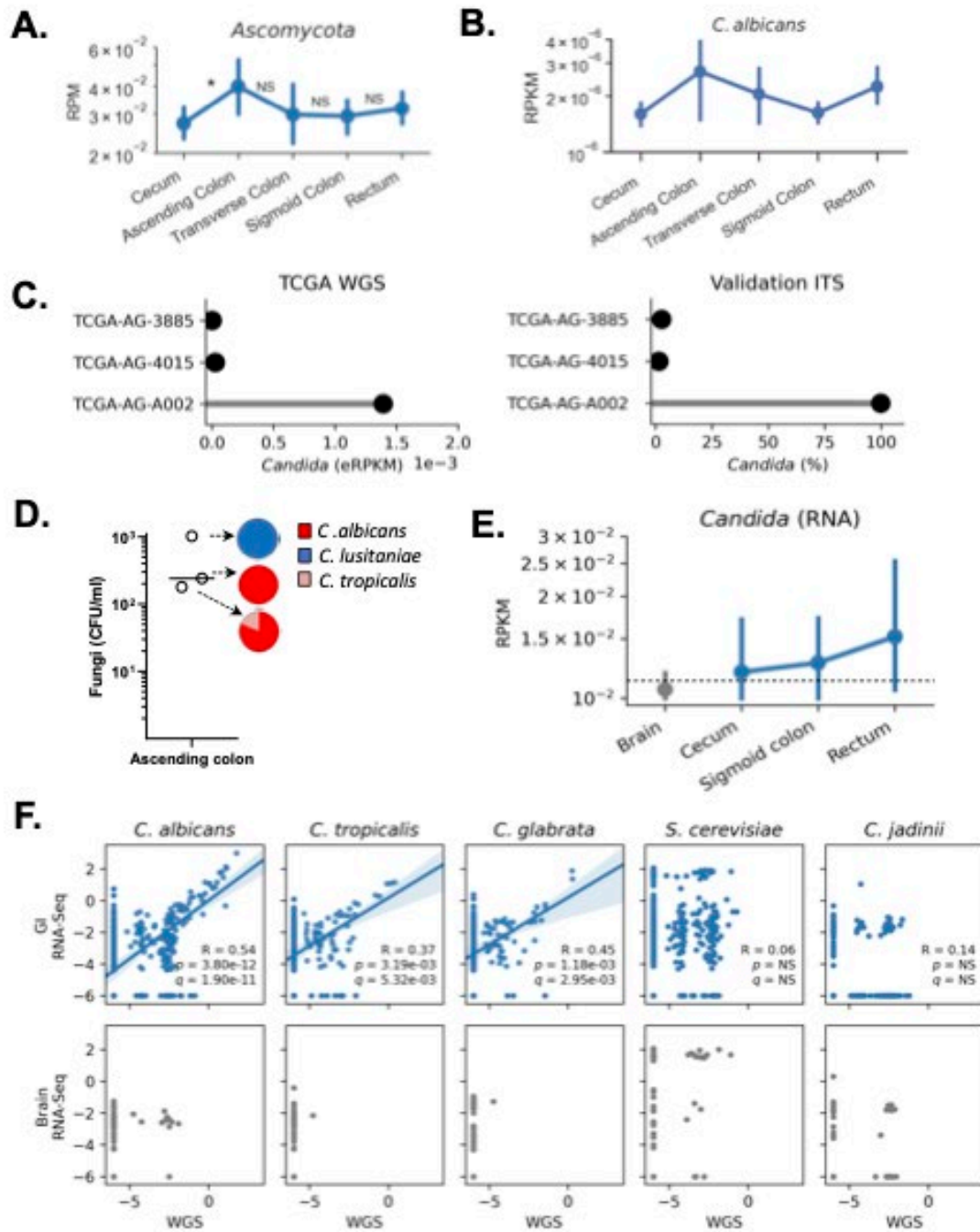


Figure 4.5: Live, transcriptionally active *Candida* species are associated with GI tumors

- (A) Spatial distribution of *Ascomycota* abundance along the colorectal tract. *Ascomycota* were most abundant in the ascending colon. Significance was calculated between adjacent tumor sites
- (B) Targeted analysis showing spatial distribution of *C. albicans* abundance by reads per kilobase of genome, per million (RPKM).
- (C) Comparison of *Candida* abundance detected in TCGA WGS data (left) and in matched tissues by independent ITS sequencing (right).
- (D) Live *C. albicans* and *C. tropicalis* are present in the mucosa of adenocarcinomas from ascending colon. Isolated fungal colonies from each individual subject were identified by MALDI-TOF, and viable fungal colony forming units (cfu) per mL of sample were determined. No live *M. globosa* or *S. cerevisiae* were isolated from these tissues.
- (E) Abundance of RNA transcripts aligning to *Candida* in brain (gray) and sites across the lower GI tract (blue) from solid tissues in the HCMI cohort.
- (F) Correlation between fungal species abundances (\log_{10} -eRPKM) determined by PathSeq analysis of TCGA WGS and RNA-seq data in GI samples (top) and brain samples (bottom) indicates that *Candida spp.* are transcriptionally active in GI tissues, while other species are not.
- (G) Live, transcriptionally active *Candida* species are associated with GI tumors \

To further examine the role of *Candida*, we next performed an analysis of fungal abundance distribution across the lower GI tract. Consistent with previous studies focused on fecal mycobiota [271, 272, 275, 335], the *Ascomycota* phylum was more prevalent in the ascending colon (Figure 4.5A, Figure S4.5). A targeted, species-level analysis determined that *C. albicans* is likely driving the abundance of *Ascomycota* in the ascending colon (Figure 4.5B).

We next sought to experimentally validate the presence of *Candida* in lower GI cancer tissues. To do so, we obtained three primary colorectal tumor samples from an original TCGA tissue provider: two of these samples was classified as *Candida*-positive (TCGA-AG-A002) and two as *Candida*-negative (TCGA-AG-4015, TCGA-AG-3885). We performed independent, ITS sequencing of these three samples and confirmed the presence of high rates of *Candida* in TCGA-AG-A002 (98.89% of reads), while *Candida* appeared to be much less abundant in TCGA-AG-4015 and TCGA-AG-3885 (<2% of reads) (Figure 4.5C).

Notably, culture-dependent analysis [311] of colorectal adenocarcinomas from a separate cohort determined that live *C. albicans*, *C. lusitaniae* and *C. tropicalis* are present in the mucosa of adenocarcinomas from ascending colon (Figure 4.5D). No live *S. cerevisiae*, *M. sympodialis* or *M. globosa* were isolated from these samples. In a third cohort from the Human Cancer Model Initiative (HCMI), we screened for the presence of *Candida* RNA in solid tumor samples, finding that the distribution of *Candida* RNA along the length of the lower GI tract (Figure 4.5E) matched the anatomical distribution of *Candida* DNA in TCGA cohort (Figure 4.5B). No HCMI solid tumor samples were available from the ascending or transverse colon. The detection of live *Candida* and *Candida* RNA in GI tumors prompted us to examine if RNA from *Candida* or other species could be detected in GI tumors profiled by TCGA. Comparing the abundance of

fungal sequences from matched tumors analyzed using both WGS and RNA-seq, we found that rates of genomic *Candida* DNA were highly correlated with the presence of *Candida* RNA transcripts (Figure 4.5F), indicating that these *Candida* species were transcriptionally active across GI tumors. In comparison, no such correlations were observed for other species, including *S. cerevisiae* and *C. jadinii* suggesting that DNA and RNA obtained from these species do not represent living fungi in these tumor tissues, consistent with our culture-dependent analysis. Together, these data demonstrate that live, transcriptionally active *Candida* species are present in tissues associated with GI tumors and that fungal DNA detected in the blood of patients with lower GI tumors may originate from the gut.

4.2.8 Targeted analysis of *Candida* and *Saccharomyces* spp.

To further evaluate the prevalence of specific fungal genera across different cancer types, we performed targeted analyses of *C. albicans*, *C. tropicalis*, and *S. cerevisiae*. This analysis revealed that *C. albicans*, *C. tropicalis*, and *S. cerevisiae* were more prevalent in GI tract tumors than breast tumors or brain tumor controls (Figure 4.6A-B).

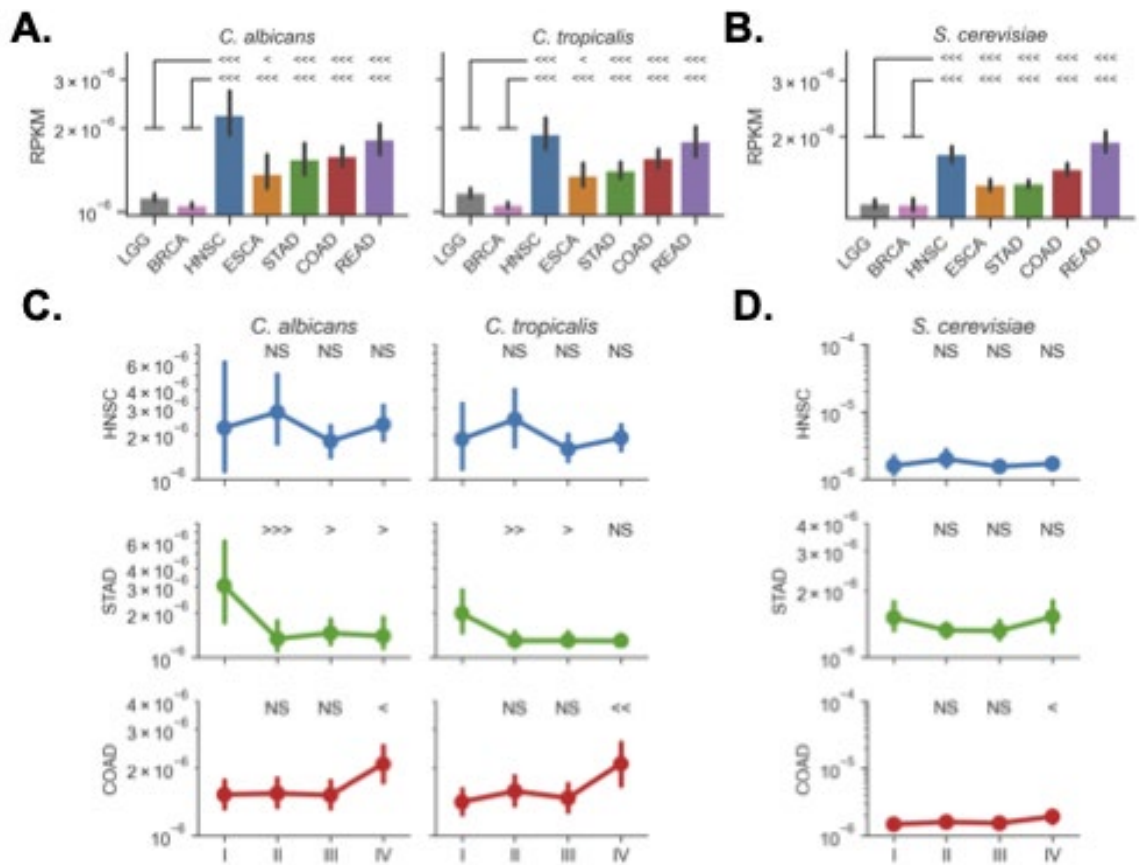


Figure 4.6: *Candida* species are present in GI cancers and high abundance is associated with early-stage stomach cancer

- (A) Targeted analysis (see “Methods”) measuring abundance (RPKM) of *C. albicans* and *C. tropicalis* across TCGA cancer types. Gastrointestinal tumors harbor higher rates of *C. albicans* and *C. tropicalis* sequences than brain and breast tissues.*
- (B) Targeted analysis shows *S. cerevisiae* is more abundant in GI tumor samples than brain and breast tissues.*
- (C) Abundance of *C. albicans* and *C. tropicalis* are elevated in stage 1 stomach cancer tumors and stage 4 colon cancer tumors. Statistical significance was calculated between stage 1 tumors and each subsequent stage.*

(D) Abundance of *S. cerevisiae* is elevated in stage 1 stomach cancer tumors and stage 4 colon cancer tumors. Statistical significance was calculated between stage 1 tumors and each subsequent stage.*

Considering our finding that *Candida*-to-*Saccharomyces* ratios may be prognostic of GI cancer outcomes (Figure 4.4C-D), we next used our targeted approach to examine associations between specific fungi and tumor stage. Consistent with our observation of *Candida*-to-*Saccharomyces* ratios, we found that *C. albicans*, *C. tropicalis*, and *S. cerevisiae* were significantly associated with stage IV colon cancer (Figure 4.6C-D). As late-stage CRC is characterized by tumor infiltration of the lymph nodes and lamina propria [236, 237], this finding suggests that *Candida* abundance may be predictive of fungal translocation to the bloodstream, a finding supported by our observation of similar fungal composition of matched blood samples (Figure 4.4E). Notably, both *C. albicans* and *C. tropicalis* were more abundant in stage I stomach cancer specifically (Figure 4.6C-D). None of the fungal species we examined were associated with a specific tumor stage in head-neck samples. These data collectively imply the presence of tumor-associated mycobiomes that may serve as prognostic markers for predicting cancer progression and patient outcomes. Furthermore, this targeted analysis indicated that increased abundance of *Candida* in late-stage, metastatic colon tumors may be directly or indirectly involved in the deregulation of genes mediating cellular adhesion (Figure 4.4B), thereby

leading to a deteriorated epithelial barrier, metastasis, and translocation of fungi from the primary tumor site into the bloodstream. Alternatively, increased abundance in late-stage colon tumors might instead be the result of deregulations in the tumor's immune system, which would allow the unhindered growth of *Candida* and other pathogens.

4.2.9 Cancer-associated mycobiota and clinical outcomes highlight predictive value of *Candida*

Having observed that higher rates of *Candida* were associated with increased expression of immune/inflammatory genes in GI cancers (Figure 4.4B-D), we sought to further explore associations between specific fungi and GI cancer types by comparing abundance of *Candida* between tumor samples and normal tissue. We found that *Candida* was significantly and uniquely enriched in stomach tumor samples compared to patient-matched normal tissue ($p = 4.23E-3$, $q = 0.026$, Figure 4.7A-B), while *Cyberlindnera* was significantly enriched in normal tissue ($p = 2.15E-5$, $q = 1.29E-4$). Notably, the same analysis determined that *Blastomyces* ($p = 8.80E-3$, $q = 0.114$) was similarly enriched in lung tumors compared to matched adjacent normal tissue (Figure S4.7A).

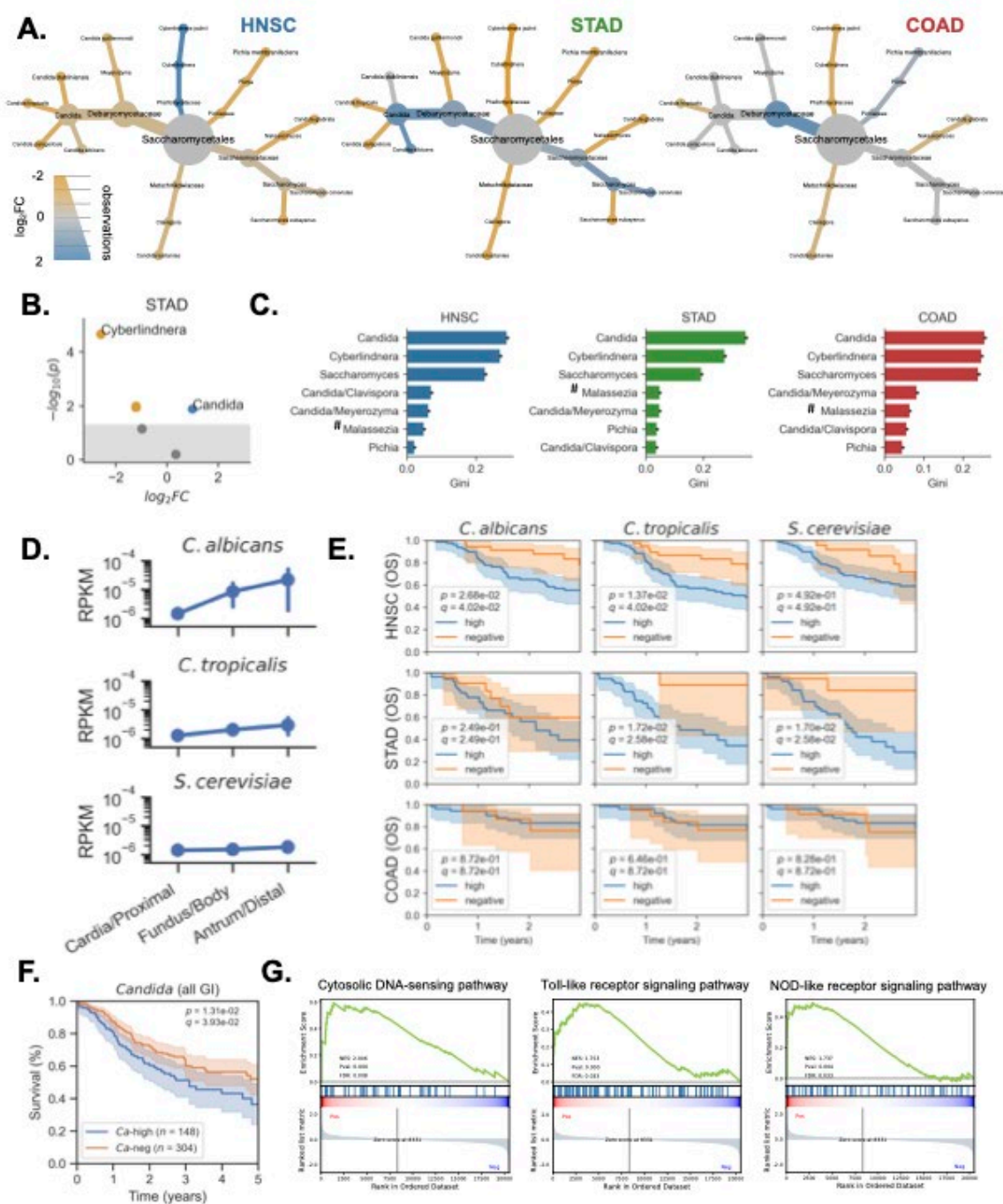


Figure 4.7: Cancer-associated fungal mycobiota and clinical outcomes highlight predictive value of *Candida*

- (A) Heat-tree depicting differential abundance of genera between tumor (blue) and matched adjacent normal tissue (yellow) in head-neck (HNSC), stomach (STAD), and colon (COAD) cancers. Across cancer types, *Candida* is enriched in tumors compared to uninvolved tissue.
- (B) Volcano plot showing differential abundance of genera between tumor (blue) and matched adjacent normal tissue (yellow) in stomach cancer. *Candida* is enriched in stomach tumors compared to matched adjacent normal tissue.
- (C) Genera identified as predictive features, used for distinguishing head-neck, stomach, and colon tumors from other tumor types. Feature importances were calculated from random forest (RF) classifiers using the gini coefficient. Site specific contaminants (#) were set to 0 prior to running the analysis and therefore some are predictive due to their absence in samples.
- (D) Targeted analysis of *Candida* spp. shows that *C. albicans* and *C. tropicalis* increases in abundance from the proximal to distal stomach, while *S. cerevisiae* abundance is relatively stable across these sites.
- (E) Survival analysis comparing outcomes for stomach cancer patients with high rates of intratumoral *C. albicans*, *C. tropicalis*, and *S. cerevisiae*, compared to patients whose blood was negative for these species. These findings suggest that *C. tropicalis* and other *Candida* spp. may be prognostic of survival in stomach cancer.
- (F) Across GI cancer types, patients with high levels of intratumoral *Candida* experience decreased survival compared to *Candida*-negative patients.
- (G) Gene-set enrichment analysis (GSEA) reveals that genes related to cytosolic DNA sensing, Toll-like receptor, and Nod-like receptor signaling are up-regulated in stomach cancers with higher rates of *Candida*.

Our analysis of TCGA GI tumor samples suggested the possibility that detection of *Candida* DNA may have potential as a prognostic biomarker. To examine this possibility, we employed a non-parametric machine learning ensemble method known

as a random forests (RF) classifier. Similar machine-learning approaches have been used previously to distinguish between cancer types based on intratumoral and circulating bacterial DNA [62]. However, RFs are particularly useful for estimating the importance of certain features for prediction. This machine learning approach found that *Candida* was by far the most important feature for distinguishing GI tumors from other cancer types, followed by *Cyberlindnera* and *Saccharomyces* (Figure 4.7C). Additional targeted analyses of *C. albicans* and *C. tropicalis* revealed that the abundance of both *Candida* species increased steadily from the proximal to distal stomach, with the lowest abundance in the cardia and the greatest abundance in the antrum (Figure 4.7D). These results mirror the colonization pattern of *H. pylori*, which preferentially infects the antrum [325].

Enrichment of *Candida* in tumor samples and its predictive power for GI cancer led us to question if *Candida* might be prognostic of disease outcomes. Using survival data from TCGA, we found that high rates of tumor-associated *C. tropicalis* DNA were significantly associated with decreased survival among stomach cancer patients ($p = 1.72E-2$; $q = 2.58E-2$) and head-neck cancers ($p = 1.37E-2$; $q = 4.02E-2$), indicating that the presence of *Candida* DNA at the tumor may be used as a non-invasive biomarker for predicting GI cancer disease outcomes (Figure 4.7E). Moreover, we observed that the presence of *Saccharomyces* spp. at the tumor site were also predictive of decreased

survival in stomach cancer, suggesting these taxa may be prognostic for multiple GI cancers (Figure 4.7E).

As the presence of *Candida* spp. appeared to be predictive of patient survival in several GI tissues, we next sought to determine if these associations extended beyond specific cancer types. To explore this possibility, we performed a pan-cancer analysis, incorporating fungal abundance and survival information from all GI cancer types, including head-neck, esophageal, stomach, colon, and rectal cancers. This analysis found that GI cancer patients with high levels of *Candida* at the tumor site had significantly decreased survival rates compared to patients who were *Candida*-negative ($p = 1.31E-2$, $q = 3.93E-2$, Figure 4.7F) while *Saccharomyces* presence at the tumor site was not associated with survival (Figure S4.6B). The associations between *Candida*, GI cancer, and reduced survival were particularly pronounced in stomach cancer, and were consistent with the results of our pathway analysis using functional KEGG pathways with GSEA, which found that the presence of *Candida* was associated with the expression of genes involved in cytosolic DNA sensing, Toll-like receptor signaling, and Nod-like receptor signaling in stomach cancers (Figure 4.7G). Together, these data not only contribute to a growing body of evidence suggesting that *Candida* contributes to GI cancer severity, but also suggest that *Candida* may serve as a promising biomarker for predicting disease outcomes.

4.3 Discussion

In this pan-cancer analysis of tumor mycobiomes, we screened NGS data from TCGA to extract and characterize the fungal composition of hundreds of tissue and blood samples from both GI and non-GI cancer types. Although fungi comprise only a fraction of the overall microbiome, our analysis demonstrates that metagenomic screens of high-throughput human sequencing data are capable of detecting tumor-associated fungal DNA. Overall, found an upper limit of about 0.1 fungal reads per million in most human tumors. Since fungal genomes are roughly 1000 times smaller than the human genome, this corresponds to roughly 1 fungal cell per 10,000 human cells. We believe that this rate is biologically plausible, given (1) the “concentration” of such fungal cells at only few barrier surfaces, and (2) recent estimates for the number of bacteria in the body that is of the same order as the number of human cells [14].

To precisely determine the fungal composition of these sample types, we applied multiple, orthogonal quality control models to identify and remove potential contaminant fungi and false-positive signals. Here, we showed that a thorough examination of genome-wide coverage depth and horizontal read distribution patterns are capable of identifying both biological contamination and false-positive assignments, respectively. This approach, in conjunction with previous metagenomic studies of publicly available NGS data [62, 72, 298], indicates that careful analysis of existing

sequencing data yields cost-effective and biologically meaningful metagenomic profiles which can be leveraged to study multi-kingdom microbe-microbe and host-microbe interactions at the cellular interface between microorganisms and the body sites they inhabit. The capacity to simultaneously profile microbial and tumor DNA should be taken into consideration when designing such experiments.

Our analysis of tumor mycobiomes revealed both pan-cancer and cancer-specific associations between tumor-associated fungi and human cancers. For example, we found that *Blastomyces* was enriched lung tumors and detected in lung cancer patients at rates far exceeding epidemiological baselines [319]. Across GI sites, we found that tumor-resident fungi cluster into *Candida*- and *Saccharomyces*-associated co-abundance groups, and that GI tumors could broadly be classified as *Ca*-type and *Sa*-type. Moreover, a trans-kingdom analysis showed that *Candida* and *Saccharomyces* are key determinants of the intratumoral bacterial communities and are each associated with distinct bacterial microbiomes in upper and lower GI cancers. We therefore hypothesize that some *Candida* and *Saccharomyces* spp. may act as “keystone taxa” [336] in the tumor microbiome, potentially driving ecological interactions and overall variation in multi-kingdom microbial composition. Such changes in tumor-associated microbial communities are likely to have significant effects on the tumor immune environment and are therefore expected to greatly influence the course of tumorigenesis and tumor

progression. Accordingly, we found that *Ca*-type tumors were associated with increased expression of IL-1 pro-inflammatory immune pathways, neutrophils, and a Type 17 immunity signatures, particularly in stomach cancer. In the lower GI tract, we found that *Ca*-type tumors were associated with a deregulation of genes involved in maintaining cellular focal adhesions and were significantly more likely to be metastatic. Further analyses of the RNA-seq data showed that *Candida* spp. was transcriptionally active in these tumors; coupled with the isolation of live *Candida* spp. in tumor samples, this raises the possibility that live *Candida* is involved in these inflammatory, pro-metastatic signatures. Additionally, we found that tumor and blood samples from the same patient harbor similar fungal compositions, particularly in lower GI cancers.

Increased tight junction permeability and loss of epithelial barrier function are common features of lower GI cancers in particular [337], and are significant risk factor for metastasis [56]. Transformation of intestinal epithelial cells to a mesenchymal-like state is encouraged by chronic inflammation [338], a process that is enhanced by highly dysbiotic, pro-inflammatory microbiota [54, 55]. In this work, we show that a *Candida*-to-*Saccharomyces* ratio was increased in late-stage colon cancers and was predictive of metastasis. As *Candida albicans* potentiates intestinal inflammation via IL-1-dependent mechanisms [311], it is reasonable to hypothesize that *Candida* contributes to inflammatory tumorigenesis in cooperation with other pro-inflammatory

microorganisms in the lower GI tract [339]. Chemotherapy and other immunosuppressive agents promote *Candida* expansion, particularly at the tumor site [339]. Additionally, inflammation has been shown to strongly promote *Candida* colonization, and *Candida* maintains this pro-inflammatory environment by itself augmenting inflammation [340]. Thus, effective prevention and management of *Candida* infections and associated inflammation might be a reasonable co-therapeutic option during cancer therapy.

The identification of varying signatures in stomach and colon cancers suggests that *Candida* may play diverging or complementary roles in each cancer type. *H. pylori* is known to promote tumorigenesis in stomach cancer [41, 325] yet is found at much higher rates in adjacent normal tissue than at the tumor site itself, since the gastric tumor microenvironment is largely inhospitable to the species' lifecycle [326]. Our analysis of stomach cancers showed that *Candida* is (1) positively correlated with *Lactobacillus spp.*, (2) anti-correlated with *H. pylori*, and (3) most abundant in the distal stomach, which *H. pylori* is also known to preferentially colonize [325]. *Lactobacillus spp.* have been shown to affect the attachment of both *H. pylori* [341] and *Candida spp.* [342]. Thus, exact ways in which *Candida spp.*, *H. pylori*, and *Lactobacillus* interact with the host both spatially and temporally during the development and progression of stomach cancer warrant further exploration.

Overall, we found that tumor-associated *Candida* was significantly associated with worse survival outcomes across GI sites. Given our findings that *Candida* is correlated with pro-inflammatory gene expression and predictive of metastasis, it is apparent that future work is needed to better understand the intricacies of *Candida* species interaction with the host during tumor development and progression. Additional mechanistic and reductionist studies may help to clarify whether tumor-associated *Candida* is driving these signatures [343]. Regardless, *Candida*'s associations with patient survival and enrichment in tumor samples compared to uninvolved tissues indicate that the identification of fungal DNA at the tumor site may provide a predictive biomarker for GI cancers.

4.4 Methods

4.4.1 Detection and quantification of mycobiota in TCGA sequencing data

The TCGA project collected biospecimens including primary tumors, normal tissue, and blood samples from cancer patients both prospectively and retrospectively until 2013. Raw TCGA sequencing data and the analyte, sample, and patient metadata (including information on tumor stage, location, metastasis, etc.) associated with each sequencing run were obtained from the NCI Genomic Data Commons (GDC) via the GDC's application programming interface (API). Raw WGS data are available from the GDC's legacy archive (<https://portal.gdc.cancer.gov/legacy-archive/>). Overall, we

analyzed data from 1,759 sequencing runs for HNSC (n = 338), ESCA (n = 143), STAD (n = 321), COAD (n = 300), READ (n = 127), BRCA (n = 230), LUSC (n = 100), and LGG (n = 200) projects from TCGA with WGS data available. From HCMI, we analyzed data from 34 sequencing runs on solid tissue samples from brain (n = 13) and lower Gi sites (n = 21).

All WGS and RNA-seq data from TCGA and HCMI were screened for fungal content using PathSeq pipeline [71, 299], which is made available as part of the Broad Institute's Genome Analysis Toolkit (GATK version 4.0.3) and relies on the Burrows-Wheeler Aligner (BWA-MEM) [344]. Prior to screening for microbial alignments, PathSeq performs multiple, iterative subtractive alignments of these previously unaligned to a host genome reference [71]. The core host reference genome used was GRCh38 (hg38). This host reference is supplemented by (1) highly variable sequences from the immunohistocompatibility complex (MHC) from the Immuno-Polymorphisms Database (IDP), (2) Cloning vector sequences from NCBI UniVec, (3) mammalian consensus repetitive sequences from RepBase, (3) a curated database of human transcripts (human v25) from Gencode, and (4) human breakpoint sequences from GenBank (KY503218, KY5808060). Reference genomes for this analysis were obtained from the PathSeq resource bundle. These files were accessed via ftp from the Broad Institute (<ftp.broadinstitute.org/bundle/beta/PathSeq/>). PathSeq was used with default

settings, except for the “minClippedReadLength” parameter, which was set to 50 for WGS and 45 for RNA-seq, since the maximum read length for the TCGA RNA-seq data is 50. All sequencing data were analyzed on a local high-performance computing (HPC) cluster with 60 compute nodes, 1,512 CPU cores, and approximately 15TB of RAM.

To isolate the endogenous fungal composition of these samples, sequencing reads from taxa at the genus and species level were normalized (1) by genome size (i.e. per kilobase of mapped fungal genome), (2) by the expected accuracy of the taxonomic assignment (i.e. weights are divided by the number of ambiguous alignments), and then (3) by the total library size (i.e. per million primary sequencing reads, regardless of alignment). These normalizations produced an “expected reads per kilobase of genome, per million primary reads” statistic (eRPKM). Kingdom- and phylum-level read counts were normalized to the library size (reads per million, RPM), as these alignments are much less prone to ambiguous assignment or significant fluctuations in genome size. Relative abundance (%) values were calculated by scaling eRPKM values, such that the sum of taxa abundances from a given taxonomic rank and sample sum to 100.

4.4.2 Quality control by removal of fungi associated with TCGA sequencing batches

To mitigate the possibility of fungal contamination in the mycobiomes we analyzed, we performed a screen to identify species and genera that showed signs of technical variation, but not biological variation. We therefore devised a two-step

prevalence-based decontamination model (See Methods) to identify and remove (1) fungal taxa whose presence was associated with specific sequencing batches and could not be explained by biological variation, and (2) samples from multi-well sequencing plates with strong evidence of contamination.

To identify contaminant taxa, we determined the prevalence of species and genera, first across each sequencing batch (plate id), then for each tumor type (TCGA sequencing project) and compared these to their expected frequencies assuming a random distribution. Specifically, expected frequency distributions for each species were calculated by multiplying the number of total number of samples in each project or sequencing plate by the species prevalence across the entire dataset; these values were compared to the observed prevalence across projects or plates. We used these observed and expected frequencies to compute p -values for a Chi-square statistic, which we adjusted for multiple comparisons using the Benjamini-Hochberg false-discovery rate correction (FDR, q -values). Species and genera that were significantly associated with sequencing batch ($q < 0.1$) but not tumor type ($q > 0.1$) were classified as potential contaminants and removed from downstream analysis. Lastly, we screened samples to determine if there were sequencing plates with significant evidence of contamination that needed to be excluded from the analysis entirely. This analysis identified a single sequencing plate (A19H), samples from which harbored fungal reads at rates that were

around five magnitudes greater than samples from different plates, independent of sample type. Overall, this analysis resulted in the identification of 35 contaminant taxa (12 genera, 23 species), and 18 of 29 contaminated sequencing runs from a single plate which were removed from downstream analysis.

4.4.3 Quality control by vertical and horizontal analyses of fungal genome coverage

To further address the possibility of contamination or false-positive alignments, we sought to characterize the genomic coverage of the species which were most frequently found in our PathSeq analysis of WGS data from TCGA. We selected any species detected in more than 5 sequencing runs (eRPKM > 0) in any of TCGA sequencing projects we analyzed (HNSC, ESCA, STAD, COAD, READ, LUSC, BRCA) that remained our precursory decontamination analysis of sequencing batches, as well as several closely related species with NCBI reference genomes available. For sequenced tumor samples from each cancer type, the human subtracted PathSeq BAM file outputs were converted back to their raw, unmapped, reads using SAMtools v1.14 [246]. Raw reads were aligned using the Burrows-Wheeler Aligner (BWA) [344] to each species' reference genome to create a new BAM containing only reads mapped to that reference. BEDTools [248] genomeCoverageBed was then used to generate coverage results with -bg flag to output statistics in bedgraph file format. Each tumor type's bedgraphs were then pooled together and their genome coverage was assessed using deepTools2 [345] bamCoverage command. Genome alignments were visualized using pyGenomeTracks [346].

We used the resulting bedgraphs to analyze the coverage depth and horizontal read distribution for each genome. Coverage depth (Vertical quality control model) was assessed by calculating the average \log_{10} -coverage per-base per-sample. We then calculated the ratio of average \log_{10} -coverage per-base per-sample between each sequencing project and brain tumor samples to estimate the fraction of reads that could be the result of contamination. To assess horizontal distribution (Horizontal quality control model) for each species and cancer type, we generated a genome-length Boolean vector indicating whether reads had aligned to each base. We then calculated the hamming distance between the vector generated for brain tissues and the vector for each cancer type to determine the base-wise horizontal similarity of alignments across each genome. For the vertical quality control model, species were classified as possible contaminants if the average \log_{10} -coverage per-base per-sample coverage for each tumor type was greater than 30% that of brain tumors. For the horizontal quality control model, species were classified as possible false-positive signals if the hamming distance to brain was less than 0.02. Species which were classified as possible contaminants or false-positive signals by either model were removed from downstream analysis.

4.4.4 Validation with TaxaTarget

We used TaxaTarget [309] to validate the presence of key species using an analysis eukaryotic marker genes. The human filtered PathSeq output BAM files from TCGA were converted to their raw, unaligned forward and reverse fastq formats using samtools. They were then screened for marker genes aligning to *Homo sapiens* to

determine the degree of contamination by human DNA, as well *Candida*, *Sacharomyces*, and *Malassezia* species to validate their presence in TCGA tumor samples.

4.4.5 Targeted analysis and quantification of *Candida* and *Saccharomyces* species of interest

We identified several species of interest that were abundant across TCGA tissue samples. To better quantify these species, we performed a targeted analysis by mapping fungal genomes to libraries putative microbial reads for each TCGA sequencing run, generated after stringent filtering of human sequences with PathSeq [71, 299].

Representative genomes for *C. albicans* (GCA_003454735.1), *C. tropicalis* (GCA_000633855.1), and *S. cerevisiae* (GCA_000146045.2) were downloaded from GenBank and mapped to these libraries using STAR [244] without allowing for spliced alignments (--alignIntronMax=1). Raw read counts for each species were then normalized by genome size and total library size as described above, to calculate an empirical reads per kilobase of genome, per million primary reads (RPKM).

4.4.6 Estimation of intra- and inter-kingdom co-abundance groups and associated gene expression signatures

It is well accepted that microbiome data should be treated as compositional, a characteristic which typically complicates robust calculation of correlations between microbiota [155, 333]. To control for compositional effects, we used SparCC [155] to estimate taxa that are frequently found together across each cancer type. This method

relies on a bootstrapping procedure to control for spurious results common in microbiome survey data. Prior to calculating correlations, we filtered out low-abundance samples and selected the 20 most abundant fungal species from each cancer type. We then ran the SparCC algorithm for 1000 iterations with default parameters to identify fungal co-abundance groups within head-neck (HNSC), stomach (STAD), and colon (COAD) tumor samples.

Our trans-kingdom analysis was used to identify associations between fungi and bacteria and was performed by comparing the decontaminated fungal compositions generated in the current work with decontaminated bacterial compositions from matched samples in TCMA [298]. To accurately quantify associations across kingdoms and control for the significant difference in their respective abundances, we applied a scaling factor to the fungal compositions in order to generate similar distributions for each kingdom and allow robust estimation of co-abundance between fungal and bacterial compositions. We then selected the most abundant fungal and bacterial taxa from each cancer type and again applied SparCC.

4.4.7 Acquisition and analysis of original TCGA tissue samples

For validation of *Candida* abundance, we obtained original, matched tissue and plasma samples from three CRC patients from Indivumed, an original TCGA tissue provider. Tumor tissues were minced, homogenized and treated with 200 U/mL lyticase

(Sigma) followed by bead beating, and processing using the Quick-DNA Fungal/Bacterial Kit (Zymo Research) as in [311]. Fungal DNA presence was validated by RT-PCR for fungal 18S and fungal ITS1-2 regions were amplified by PCR using primers with sample barcodes and sequencing adaptors.

Fungal primers: ITS1F-CTTGGTCATTTAGAGGAAGTAA

ITS2R-GCTGCGTTCTTCATCGATGC

Forward overhang: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-[locus-specific sequence]

Reverse overhang: 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-[locus-specific sequence]

ITS amplicons were generated with 35 cycles using Invitrogen AccuPrime PCR reagents (Carlsbad). Amplicons were then used in the second PCR reaction, using Illumina Nextera XT v2 (Illumina) barcoded primers to uniquely index each sample. 2x300 paired-end sequencing was then performed on the Illumina MiSeq (Illumina). DNA was amplified using the following PCR protocol: Initial denaturation at 94°C for 10 min, followed by 40 cycles of denaturation at 94°C for 30 s, annealing at 55°C for 30 s, and elongation at 72°C for 2 min, followed by an elongation step at 72°C for 30 min. All libraries were subjected to quality control using DNA 1000 Bioanalyzer (Agilent), and Qubit (Life Technologies) to validate and quantify library construction prior to

preparing a Paired-End flow cell. Samples were randomly divided among flow cells to minimize sequencing bias. Clonal bridge amplification (Illumina) was performed using a cBot (Illumina). 2 x 250 bp sequencing-by-synthesis was performed on Illumina MiSeq platform (Illumina).

4.4.8 Quantification, isolation, and characterization of live fungi in primary colorectal tumor samples

Adenocarcinoma-associated tissues were collected from ascending colon surgical resections that were then weighed, minced, homogenized, diluted in sterile PBS and plated onto Sabouraud dextrose agar (SDA) and modified Dixon media (mDixon with glycerol monostearate), and inhibitory mold agar (Hardy Diagnostics), all supplemented with both penicillin/streptomycin (Sigma), inhibitory mold agar (Hardy Diagnostics) and modified Dixon broth with glycerol monostearate. SDA plates were incubated at 37°C for 48 hours. Inhibitory mold agar plates and modified Dixon media were incubated at 30°C for up to a week. Isolated fungal colonies from each individual subject were identified by matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometer.

4.4.9 Identification of *Candida*- and *Saccharomyces*-type TCGA tumor samples and associated signatures

To identify *Candida*- and *Saccharomyces*-associated tumors, we calculated a log₂ *Candida*-to- *Saccharomyces* abundance ratio ($\log_2(C/S)$) across all tumor samples for which

either genus was detected. Tumors were classified as *Ca*-type or *Sa*-type if they had a $\log_2(C/S)$ value above 1 or below -1, respectively, i.e. samples for which neither genus was detected at more than twice the rate of the other were excluded. To test associations between gene expression and the presence of *Candida* and *Saccharomyces*, we performed differential gene expression analysis using batch-normalized gene expression data from the PanCanAtlas publication page (<https://gdc.cancer.gov/about-data/publications/pancanatlas>). For each cancer type, we calculated \log_2 -fold changes (\log_2FC) in gene expression between tumors that were negative for *Candida* or *Saccharomyces* (eRPKM < 1E-6) and tumors which were high in *Candida* or *Saccharomyces* (eRPKM > 1E-6). All taxonomic abundance profiles were collapsed to the sample level using the geometric mean of taxon abundances across the available tumor sequencing data for each tumor sample. We then estimated the significance of gene expression changes using Student's independent two-sample *t*-test. The differential gene expression values generated by this analysis were then used to perform GSEA [234] and analyze gene expression pathways enriched in *Candida*- and *Saccharomyces*-associated cancers based on gene lists obtained from MSigDB v7.1. Using pre-ranked differential gene expression values, we ran GSEA for 1000 iterations to identify KEGG biological pathways [347] that were enriched in *Candida*- and *Saccharomyces*-associated tumors from each cancer type.

To compare rates of metastasis in *Ca*- and *Sa*-type tumors, we used TNM-stage classifications of each TCGA tumor sample to determine metastatic (M1) or non-metastatic (M0) status. Samples for which no metastatic information was available (MX) were excluded. We then generated a contingency table for each cancer type comparing metastatic status (M0/M1) and tumor mycobiome classification (*Ca*-type v.s. *Sa*-type) and used Fisher's exact test to determine whether *Ca*-type or *Sa*-type tumors were more likely to be metastatic.

4.4.10 Differential abundance analysis between tumor and adjacent normal tissue

Associations between fungal genera and sample type (tumor vs. matched adjacent normal tissue) were calculated in R, using a custom paired analysis function written for metacoder [256]. For each cancer type we analyzed we selected the 20 most abundant taxa, provided they were present in at least 30 samples overall. Such filters help to remove low-abundance and low-prevalence taxa which frequently have small means and large coefficients of variation, contributing unnecessary noise for downstream differential abundance comparisons. After adding pseudocounts, we calculated the relative abundance of fungi for each sequencing run. Across all patients with matched tumor and normal tissue, we then calculated the median ratio of each taxon's relative abundance values in tumor samples compared to matched adjacent normal tissue. Significance values were calculated for \log_2 median ratios between

transformed relative abundance values, using Wilcoxon's rank-sums test. Taxa with significant p -values ($p < 0.05$) were selected for downstream analysis.

4.4.11 Survival analysis

We performed our survival analyses using the log-rank test, as implemented by the lifelines survival analysis python package (Davidson-Pilon et al., 2020). Data on TCGA patient survival were collected from the PanCanAtlas' clinical follow-up data [233]. This analysis was performed at both the species and genus level. For the species-level analysis, we used normalized fungal abundances from our targeted analysis (RPKM for *C. albicans*, *C. tropicalis*, and *S. cerevisiae*). For each species of interest and cancer type, we compared survival between patients whose tumors did not harbor the species ("negative"; 0th percentile) with patients whose tumors were abundant in the species ("high"; top 50th percentile). The genus-level analysis was performed using fungal abundances determined by our PathSeq analysis (eRPKM for *Candida* and *Saccharomyces*) and used the same set of criteria for assigning patients as "negative" or "high" as the differential gene expression analysis. All taxonomic abundances were collapsed to the patient level using the geometric mean of taxon abundances across the available tumor sequencing data for each patient.

4.4.12 Random forest classification of cancer types using fungal compositions of tumor and blood samples

To identify taxa that are predictive of cancer location, we used a decision-tree based ensemble machine learning method known as random forest classifiers [348], as implemented by the python package sklearn [349]. A separate classifier model was trained on the mycobacterial compositions of tumor samples from seven TCGA cancer types (HNSC, ESCA, STAD, COAD, READ, LUSC, and BRCA). For each cancer type, we implemented a one-versus-all classification strategy which sought to identify genera capable of distinguishing a given cancer type (e.g. stomach tumors) from all others (e.g. non-stomach tumors). Prior to classification, taxa that were detected in fewer than 1% of samples were removed. Species abundances were log-normalized after the addition of a pseudocount to achieve a gaussian distribution. For each classifier, a forest of 400 estimators was used, with a maximum depth of 30 features per tree, and a minimum of 5 samples per split. Default values were used for all other hyperparameters. To bootstrap the estimation of feature importances, we used a repeated, stratified cross-fold cross validation strategy with 10 folds and 10 repeats. Feature importances were estimated by averaging Gini impurity measures for each of the 100 resulting sub-models.

5 Taxonomic Set Enrichment Analysis: A Curated Database and Analytical Toolkit for Interpreting Metagenomic Data

5.1 Introduction

Falling costs and improvements in DNA sequencing have ushered in a renaissance in microbiome research. Whereas traditional microbiology once relied on cultivating and characterizing microbial cultures, it is now possible to rapidly and cost-effectively characterize entire microbial communities using high-throughput sequencing, revealing enormous biodiversity among both biotic and abiotic microbiomes at unprecedented scale and resolution. Both amplicon sequencing methods as well as shotgun metagenomics have dramatically expanded universe of known microorganisms [11, 12], allowing tremendous insight into the ways in which microbial communities interact with the diverse array of ecosystems they inhabit. Simultaneously, the rapid accumulation of microbiome sequencing datasets and the growing number of species they are capable of identifying have compounded the challenges of interpreting them, as gathering insights across hundreds or thousands of individual species becomes untenable.

In analyses of large gene expression datasets, gene set enrichment analysis (GSEA) is a step performed almost universally. This powerful, knowledge-based method has allowed considerable insight into the biological pathways and processes

underlying shifts in gene expression [234]. Microbe set enrichment analysis (MSEA) represents an analogous class of analytical tools which have been developed to help interpret microbiome profiling data [70, 255, 350]. However, MSEA have not seen the same degree of widespread adoption as GSEA and is rarely performed within metagenomic data analysis pipelines, due to the lack of a comprehensive microbial traits reference database, limitations in the available statistical methodology, and the inaccessibility of software implementations for these analyses.

Although considerable efforts have been made to understand the human microbiome, much of the data describing it remain highly distributed [234]. The sum total of humanity's knowledge of microorganisms exists in a vast sea of databases, publications, and human capital. Collecting biological knowledge into accessible formats thus requires the cooperatives efforts of numerous researchers and curators. To that end, Gene Ontology (GO) [351, 352] consortium, and the curators maintaining the Kyoto Encyclopedia of Genes and Genomes (KEGG) [347, 353], WikiPathways [354], Reactome [355] and the Molecular Signatures Database (MSigDB) [356] have done extraordinary work to catalogue and systematize our collective knowledge of biological pathways and processes. However, tools and resources for the interpretation of metagenomic analyses have lagged behind. This can be attributed in part to the fact that metagenomes are (by definition) more heterogeneous and structurally diverse: the development of sequencing

technologies, statistical methods, and analytical tools which are sufficiently generalizable poses a significant challenge. For example, the genomes of microbial prokaryotes, eukaryotes, and archaea vary widely in size, structure, chromosome count, taxonomy, and GC content, meaning that sequencing technologies do not necessary capture metagenomes at rates proportional to each organisms' presence in a biological sample. Indeed, there is considerable need for the improvement of statistical and methodological approaches for accurately characterizing microbial communities, as well for the interpretation of large metagenomics datasets.

Here, we present Taxonomic Set Enrichment Analysis (TaxSEA), a knowledge-based statistical analysis toolkit for interpreting the results of metagenomics experiments. TaxSEA is accompanied by a massive, curated taxonomic annotations database, which is organized using a hierarchically-organized ontology of microbial traits. The TaxSEA software package (<https://github.com/abdohlman/TaxSEA>) provides methods for performing MSEA from both discrete and continuous taxonomic queries, as well as for easy visualization and summary of the results. Additionally, we have developed an interactive website to allow users to quickly and conveniently analyze their own metagenomic data without downloading software.

To establish the TaxSEA database, we began by establishing a hierarchical ontology that categorizes microbial annotations into microbe-intrinsic traits (e.g.

molecular, cellular) and microbe-extrinsic traits (e.g. biotic, abiotic) and allowed this to guide our meta-analysis. We screened microbial annotations data in a countless number of formats, ranging from CSV files to APIs to relational database systems. We meticulously parsed through these resources to extract, annotate, and organize microbe-trait associations into a common standardized format, which relates microbial traits to specific taxa, defined by their NCBI taxonomy IDs which remains stable even if species are re-named or re-classified. Altogether, the TaxSEA database contains more than 14 million trait-microbe relationships, spanning a total of 25,504 unique features describing and 122,527 taxa. Such annotations map bacterial and archaeal species to features such as size, shape, morphology, metabolism, gene expression, as well as host-associated features like body site, disease associations, and antibiotic sensitivity, among others.

As a proof-of-principle, we test TaxSEA on several existing datasets. Analyzing differential abundance data from the Human Microbiome Project, we found that TaxSEA was able to accurately identify body site and associations with Irritable Bowel Disease (IBD) using differential abundance data mapped to a trait-microbe reference generated from independent cohorts. Additionally, we applied TaxSEA to identify common features among species associated with several conditions for which the microbiome is believed to play a role, including (1) *C. difficile* infection as compared to asymptomatic carriers, (2) response to PD-1 immunotherapy in melanoma as compared

to non-responders, and (3) colorectal tumors as compared to adjacent normal tissue.

Finally, we developed a software package and an interactive website which will allow researchers to perform TaxSEA analyses of their own data.

5.2 Methods

5.2.1 Continuous and discrete TaxSEA

Broadly speaking, enrichment analysis methodologies can be categorized into discrete enrichment analysis (DEA) or continuous enrichment analysis (CEA) [357].

Statistical approaches for DEA rely on set operations, usually by identifying a pre-selected subset or “query set” containing biological entities of interest and determining its overlap with “reference sets”, which include entities associated with a given

biological process or phenomena, given a predefined “background” of taxa. A statistical test is then used to determine if there is more significant overlap between the query set and the reference set than is expected by chance; these analyses usually incorporate Fisher’s exact test, chi-square tests, binomial tests, or hypergeometric tests [70, 357-359].

By contrast, query sets for CEA are not generated by pre-selection and instead make use quantitative information by analyzing a vector input: all entities in the space are associated with a positive or negative numerical value (often \log_2 fold change) signifying its association with given experimental condition. Statistical methods for CEA range

broadly, but often rely on parametric approaches to quantifying enrichment, such as permutation analysis and Kolmogorov-Smirnov statistics[234, 357].

To allow for the most flexible application, we implement methods for performing both DEA and CEA analyses using TaxSEA. It is appropriate to use DEA when there is a finite set of microorganisms of interest, particularly when each member of the set is as important as the others. Users provide a query set of taxa they are interested in (e.g., a list of species found to be associated with IBD) as well as background of the total set of taxa they included in their original analysis (e.g. the set of taxa they detected in the metagenomic screen, or the set of taxa included in the reference genome). Results are returned in a table, showing overlap with various reference sets, p-values calculated using Fisher's exact test, as well as false-discovery correction. It is appropriate to use CEA when a differential abundance analysis has been performed between two conditions, producing a list of taxa matched to numerical values describing their association with a given condition (e.g. \log_2 fold change in tumor compared to normal tissue). Any strategy for calculating differential abundance can be used, however the resulting data should be roughly normally distributed with the sign of the values indicating the direction of the association; a value of zero should represent "no change" between phenotypes. The methodology for CEA implemented in TaxSEA is borrowed from software developed for GSEA [234], so as to leverage methodology and data

visualizations techniques that are already familiar to the genomics research community. Like DEA, results from CEA are returned in a table, showing overlap with various reference sets, as well as p-values determined by monte-carlo sampling by random permutation of taxa.

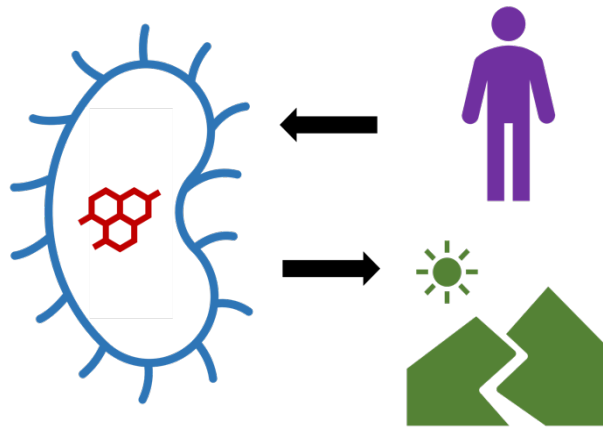


Figure 5.1: Cartoon depicting the organization of the TaxSEA database

Microbe-intrinsic traits (left) are composed of molecular (red) and cellular (blue), while microbe-extrinsic traits (right) are divided into biotic (purple) and abiotic (green) traits

5.2.2 An ontology for microbial traits

To develop our microbial traits database, we sought to establish organizing principles for categorizing of microbial traits. The resulting hierarchical ontology is intended to allow users to focus their use of TaxSEA to their desired research question. Such ontologies have been useful for previous implementations of enrichment analysis

methodology. For example, Gene Ontology (GO) [351, 352] divides genetic traits into “molecular functions”, “biological processes”, and “cellular components”.

Table 5.1: Hierarchy of microbial traits and associated sources

Level 1	Level 2	Level 3	Feature	Source
Microbe-intrinsic traits	Molecular	Genomic	Genome features	BAP
			Phylogeny	NCBI
		Transcriptomic	Signal domains	MIST, Pfam, TIGRFAM
			Gene expression	COG
		Proteomic	Protein function	COG
			Protein expression	COG
	Metabolic	Biochemical pathways	BAP, COG, MACADAM	
		Carbon substrates	BAP, NJS	
	Cellular	Lifecycle	Metabolic class	BAP, MicrobiomeAnalyst, NJS
			Growth rate	BAP
			Sporulation	BAP
		Physiology	Morphology	BAP
Geometry			BAP	
Motility			BAP	
Gram status	BAP			
Microbe-extrinsic traits	Biotic	Public health	Host disease	MicrobiomeAnalyst, Micropattern, Disbiome, iHMP, CDC, NJS, iHMP
			Host health	MicrobiomeAnalyst
			Pregnancy	iHMP
		Therapeutic	Medication	MicrobiomeAnalyst
			Antibiotic sensitivity	Wellington ICU drug manual
			Chemical perturbation	MicrobiomeAnalyst
			Immunotherapy	MicrobiomeAnalyst
		Demographic	Host geography	MicrobiomeAnalyst, AGP
			Diet	MicrobiomeAnalyst, AGP
			Lifestyle	MicrobiomeAnalyst, AGP, HMP
		Physiological	Body site	HMP
			Host genetics	MicrobiomeAnalyst, PubMed
	Enterotype		Bork2011	
	Abiotic	Ecological	Habitat	AGP
			Generalist/Specialist	Sriswasdi2017
		Biogeochemical	Growth conditions	BAP
	Oxygen tolerance		BAP, NJS	

Our microbial traits database is organized using a three-level hierarchical ontology (Figure 5.1, Table 5.1). At the highest level, we broadly classify microbial traits as “intrinsic” or “extrinsic”. In this case, microbe-intrinsic traits refer to a microbe’s natural or essential qualities (e.g. gram status), while microbe-extrinsic traits describe how the microbe interacts with its external world (e.g. oxygen preference). “Microbe-

intrinsic" traits are further divided into "molecular" and "cellular" traits, depending on whether they relate to the organism's molecular expression or the organism's cellular state. "Microbe extrinsic" traits are further divided into "biotic" and "abiotic" traits, which describe the ways in which a microbe interacts with its living and non-living environment, respectively. At the third and final level, traits are from each of the four classifications (intrinsic-cellular, intrinsic-molecular, extrinsic-biotic, extrinsic-abiotic) are grouped into general categories (e.g. genomic, ecological, etc.).

Identification of studies via databases and registers

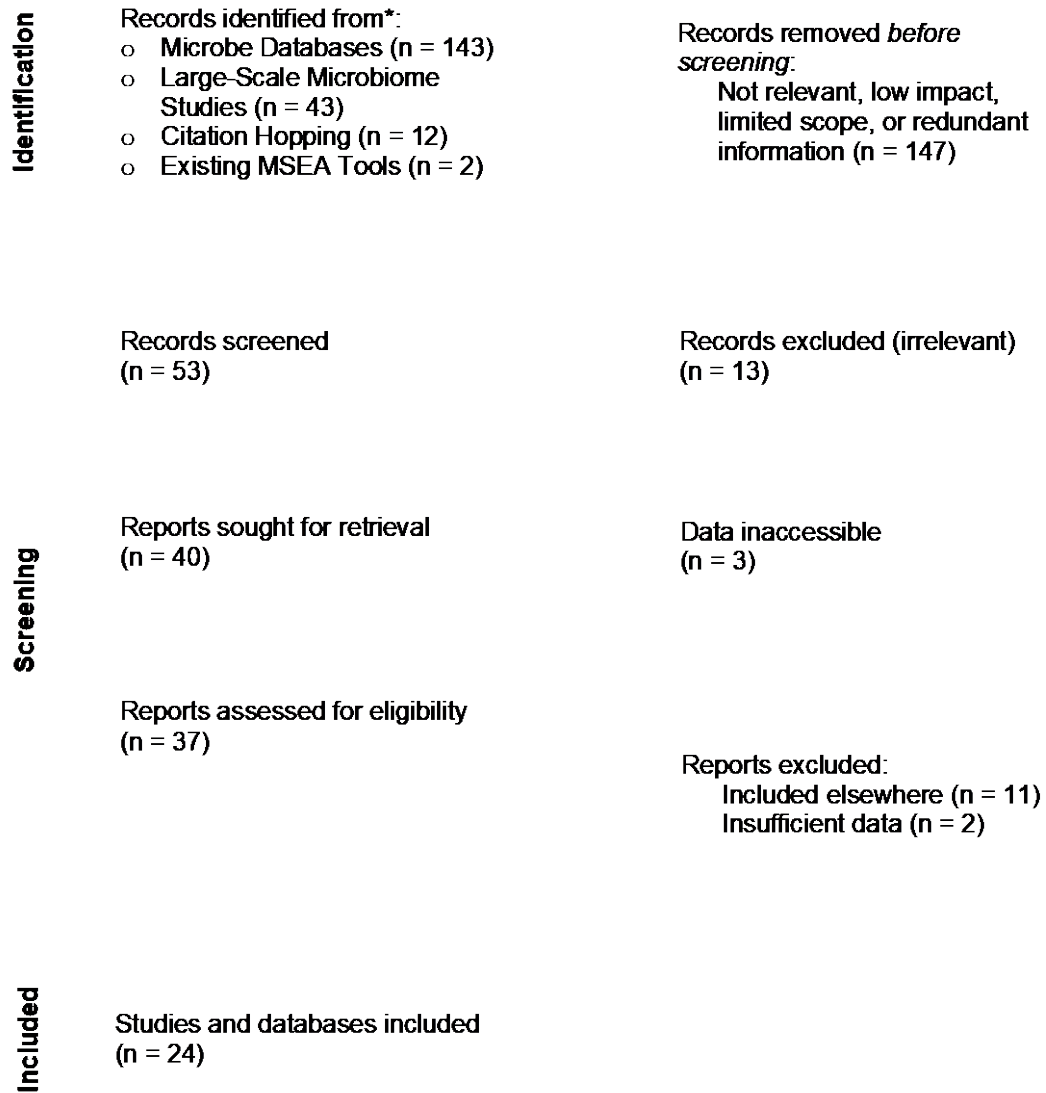


Figure 5.2: PRISMA diagram depicting the selection process for including microbiome resources in the TaxSEA database

5.2.3 Creation of a microbial traits database

To establish a comprehensive database to map microbiota to various traits describing them, we used the PRISMA framework [360] to perform our meta-analysis of microbiome trait annotations. In total, identified 143 existing databases [361], 43 large-scale microbiome studies, 2 existing MSEA databases, and 12 additional publications (Figure 5.1) [362]. There were several criteria for inclusion:

- (1) The resource was relevant, high impact, or had broad scope and could be characterized into the ontology described in Table 5.1.
- (2) The resource provided microbial annotations in an accessible format from which trait-microbe associations could be readily identified.
- (3) The resource was non-redundant, i.e. it was not already included as a subset of another resource we screened.
- (4) The resource contained sufficient microbial features and taxa to link them to.

This screen resulted in the identification of 22 total sources. These include large, well-known microbiome sequencing and annotations initiatives, including the Human Microbiome Project (HMP), The American Gut Project (AGP), the Earth Microbiomes Project (EMP), and Clusters of Orthologous Groups (COG), as well as several large-scale microbiome studies from individual publications.

To remain consistent with analytical tools and formats from gene-set enrichment methodology, we adopted the Gene Matrix Transposed (“GMT”) file format to standardize the TaxSEA database. The matrix which this file format refers to relates features (rows) to genes (columns) and is populated by binary values indicating whether each gene has a particular feature. GMTs are “transposed” to increase human-legibility and limit file size, they are generally formatted as a text file with each line containing the feature name, followed by a comma-separated list of entities associated with it (e.g. genes, taxa). For the TaxSEA database of GMTs, we adapt this format: each line contains a uniquely-assigned feature identifier, a formatted feature name, and comma-separated list of NCBI taxonomy IDs referring to the set of taxa described by the given feature. The NCBI taxonomy IDs are used instead of latin names since taxa are frequently re-named or re-classified, while the taxonomy IDs remain stable. However, TaxSEA also provides a dictionary for converting to NCBI taxonomy IDS from species names as well as several other taxa identifier formats.

5.2.4 An interactive website for interpreting microbiome data

To make TaxSEA as accessible as possible, we also developed an interactive website which will allow users to conveniently run their own analyses without the need to download software. Users can use this portal to perform both DEA by uploading differential abundance results, or CEA by uploading a list of taxa of interest. Once a list

is submitted, the website performs the analysis and renders interactive bar plots depicting the ten most enriched features for each trait category and source.

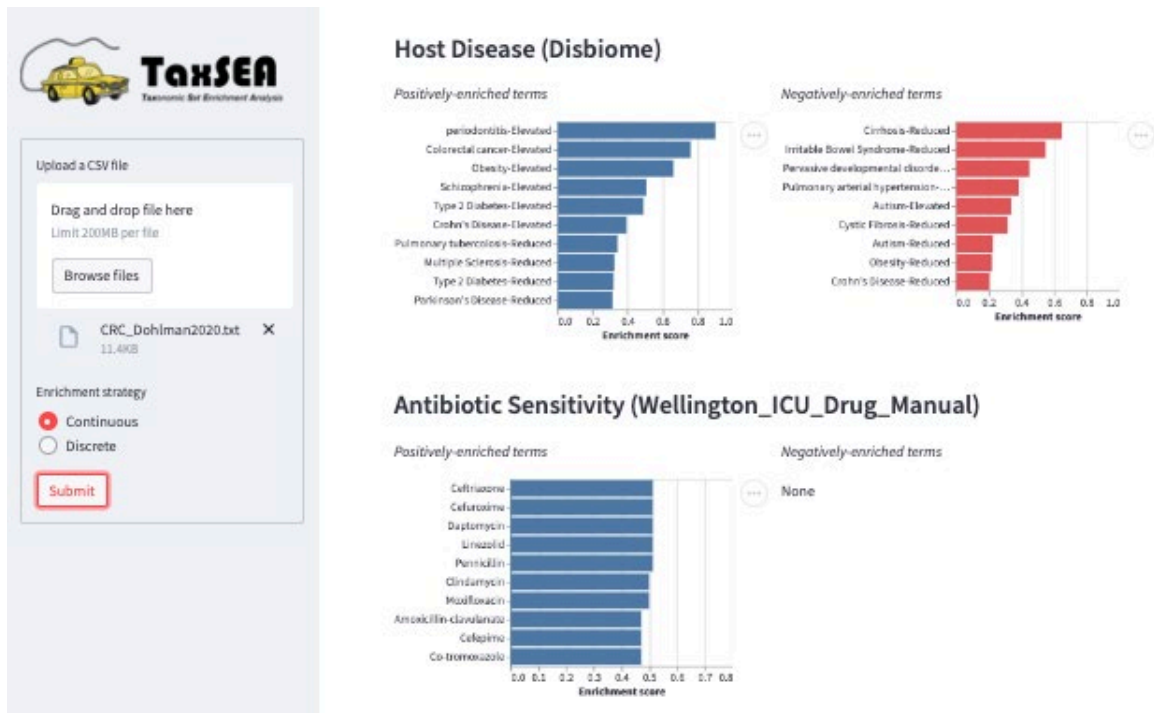


Figure 5.3: Screenshot of the TaxSEA website

5.3 Results

5.3.1 TaxSEA accurately identifies human body sites

To confirm that TaxSEA methodology works as expected, we sought to benchmark its performance using well-known microbe-trait relationships. We acquired shotgun metagenomics data from The Human Microbiome Project (HMP), including

stool and swab samples from the gut (n = 553), skin (n = 309), vagina (n = 234), and oral cavity (n = 1,259). We then used DESeq2 [363] to analyze the differential abundance between each body site in four one-versus-all (OVA) analyses. For species significantly associated with each body site ($p < 0.05$), we then performed a discrete TaxSEA analysis.

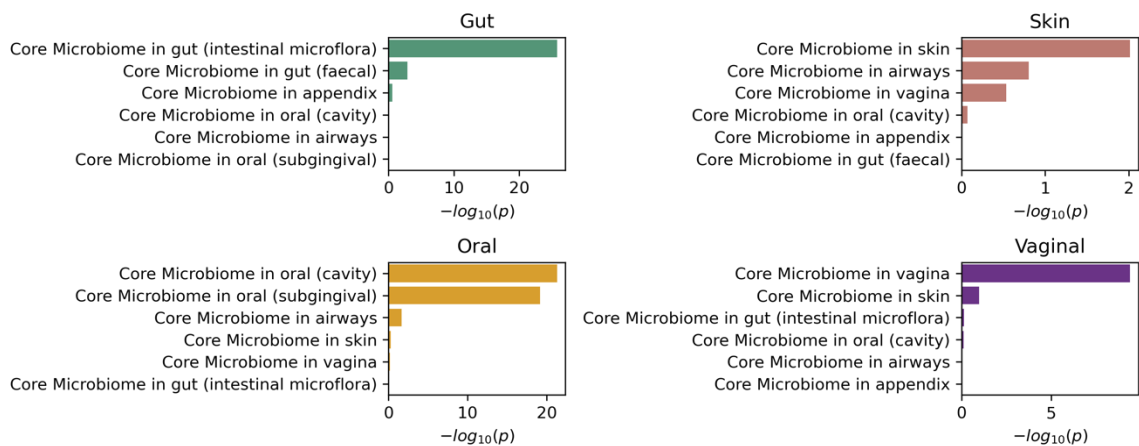


Figure 5.4: TaxSEA correctly identifies host body sites from an independent reference

We analyzed the HMP body site results using an independent reference of microbe-body site associations generated from the Genomes Online Database (GOLD) [364], obtained via MicrobiomeAnalyst [350]. Despite being a reference from an entirely different resource, TaxSEA accurately predicted the body site for each sample type: for each analysis, the correct body site was by far the most statistically significant feature identified (Figure 5.4). For the gut, the most significant feature was “Core microbiome in gut (intestinal microfloral)” ($q = 1.41E-25$), followed by “Core Microbiome in gut

(faecal)" ($q = 5.86E-3$). For the skin, "Core microbiome in skin" ($q = 0.087$), for the oral cavity "Core microbiome in oral (cavity)" ($q = 4.38E-21$), for the vagina "Core Microbiome in vagina" ($q = 3.50E-9$) were the most significant features. Together these results demonstrate that TaxSEA and the TaxSEA reference database are together capable of providing biologically accurate, statistically significant results which can be used to interpret the results of metagenomic analyses.

5.3.2 Microbiome features associated with colorectal tumors

A growing body of evidence demonstrates that colorectal cancer (CRC) tumors harbor an intratumoral microbiome which can influence cancer development, progression, prognosis, and response to therapy [28, 29, 36, 39]. Therefore, we sought to identify common features of microbiota associated with colorectal tumors. Using decontaminated microbial profiles from The Cancer Microbiome Atlas (TCMA) [36], we performed a paired differential abundance analysis between tumor samples and matched adjacent normal tissue. This analysis showed that several *Fusobacterium spp.* were significantly enriched CRC tumors, the most significant being *F. nucleatum* ($p = 1.82E-3$) which has previously been implicated in intestinal tumorigenesis in several independent reports [28, 39, 44].

Using the results from our differential abundance analysis, we performed a continuous TaxSEA analysis of the \log_2 fold change values for each species. To validate

that this analysis could accurately identify disease associations, we examined associations with a trait-microbe reference generated from the Disbiome database [69], a curated set of known associations between microbiota and disease. This analysis found that the two most significant disease associations identified were “Colorectal cancer (Elevated)” and “Periodontitis (Elevated)” (Figure 5.5A-B). Encouraged that TaxSEA correctly identified the disease type as CRC based on differential abundance, we were also fascinated to also see significant enrichment for periodontitis-associated bacteria in the tumors of CRC patients. Periodontitis is an advanced form a periodontal disease, a gum disease that is characterized by inflammation caused by dysbiosis of the normal oral microbiome [365, 366]. Epidemiological studies have found that periodontal disease is a significant risk factor for CRC, increasing the risk by 50-80% [367, 368].

These associations prompted us to further examine the microbiota associated with periodontitis. We found that several periodontitis-associated taxa, including *Paraprevotella*, *Prevotella*, and *Porphyomonas* appeared to increase with CRC tumor stage (Figure 5.5C). Additionally, using patient survival data from the TCGA PanCanAtlas, we compared the survival of patients with high rates of these species (top quartile) to those with low rates (bottom quartile), and found that *Prevotella* ($p = 6.65E-2$) and particularly *Paraprevotella* ($p = 5.65E-4$) appeared to be significant predictors of reduced overall survival in CRC patients (Figure 5.5D).

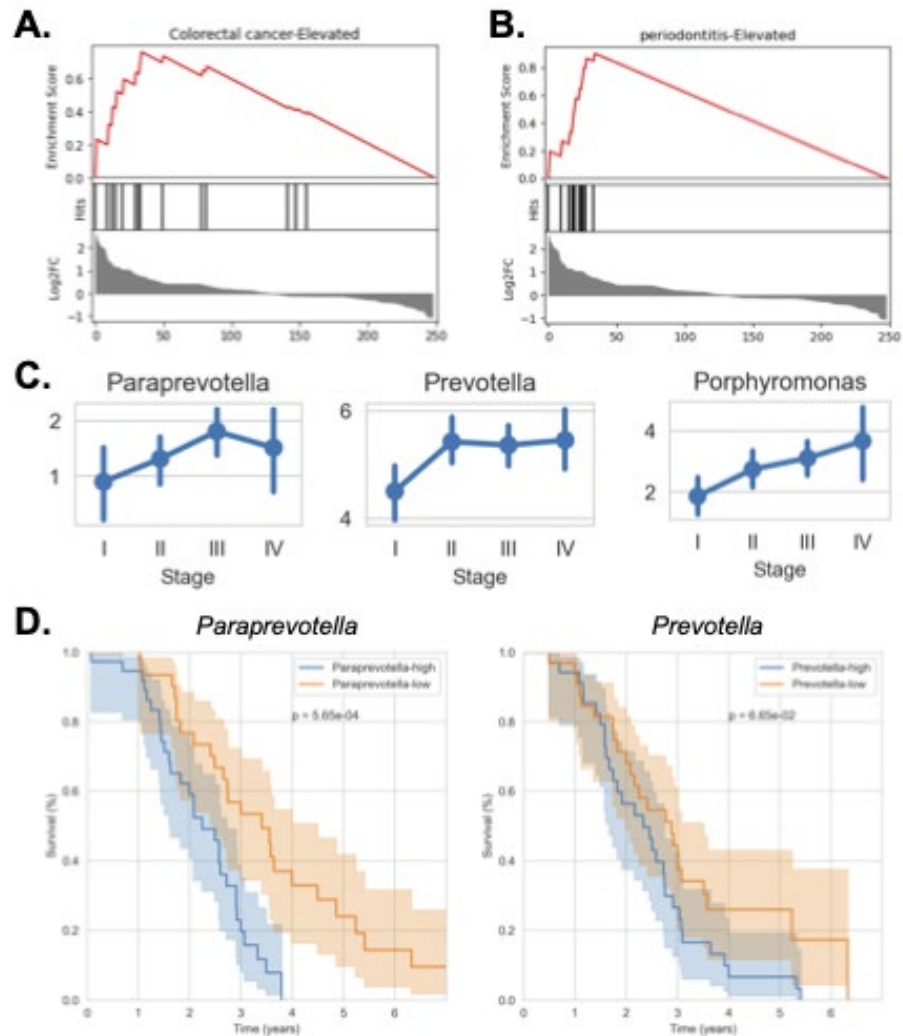


Figure 5.5: TaxSEA links colorectal cancer to periodontitis.

- (A) TaxSEA correctly identifies the disease association from a differential abundance analysis comparing the microbiomes of CRC tumors with adjacent matched controls using data from TCMA
- (B) Microbiota associated with periodontitis are significantly enriched in CRC tumors
- (C) Microbiota associated with periodontitis are enriched in late-stage CRC
- (D) Microbiota associated with periodontitis are predictive of decreased survival

Interestingly, additional analysis of CRC-associated microbiomes found positive enrichment for species known to produce ammonia and negative enrichment for species known to consume ammonia (Figure 5.6A), suggesting diverging metabolic preferences of species associated with tumor samples compared to matched uninvolved tissue, respectively. A role for ammonia has previously been identified in both CRC [369-371] and periodontal disease [372], while urea cycle dysregulation is involved in many different cancers and promotes mutation by enhancing pyrimidine synthesis and nitrogen availability [373].

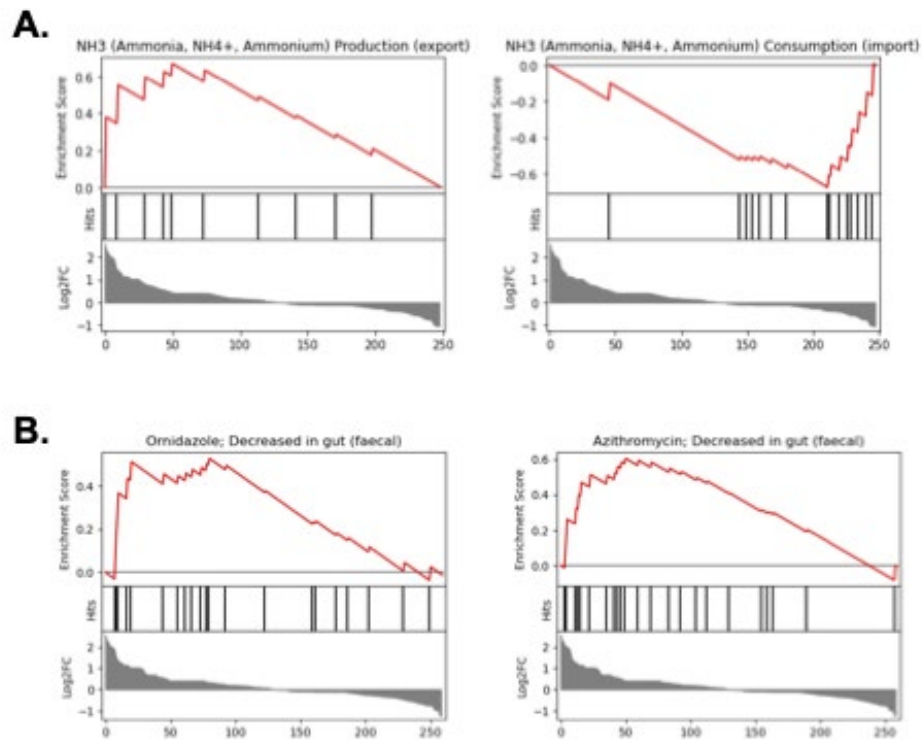


Figure 5.6: TaxSEA identifies metabolic functions and potential treatments for CRC microbiomes

- (A) **Ammonia producers and ammonia consumers are respectively enriched in tumor and matched adjacent normal tissue**
- (B) **Ornidazole and azithromycin may be used to eliminate tumor-associated microbiota**

Finally, we analyzed associations between CRC-enriched microbiota and microbial changes known to occur response to chemical perturbations of the host, including by antibiotics. This analysis found that two antibiotics, ornidazole and azithromycin were likely to target tumor-associated microbiota, demonstrating TaxSEA's ability to identify potential therapeutic options for CRC and other diseases.

5.4 Discussion

Statistical tools for analyses of the human genome and transcriptome have become well established as the result of extraordinary efforts and funding initiatives, which have created create massive, public resources for understanding the human genome and gene regulation. Since the HGP completed the first draft of the human genome, a suite of technologies, tools, and resources have made NGS analysis of the human genome and human gene expression relatively inexpensive and analyses of these data straightforward. Databases like Gene Ontology [351, 352], KEGG [347, 353], WikiPathways [354], and Reactome [355] have allowed the evaluate the available literature and assign gene classifications, determine gene interactions, and map out their

roles in various biological pathways. Together, these initiatives allowed the development of a broad ecosystem of computational models, predictive algorithms, machine learning methodology for understanding the human genome and human gene regulation. Thus, TaxSEA helps to carry this tradition into the metagenomic space

Methodological categorization of life on earth has been a centerpiece of biological research since antiquity [374]. Since the discovery of discovery of nucleic acids in the late 19th century, it has been possible rapidly and precisely identify organisms using their genetic code using high-throughput sequencing. To make sense of exponentially growing number of metagenomic datasets, there is growing demand for tools and resources to process and interpret them. To address such issues, we developed TaxSEA, a statistical tool for microbe set enrichment analysis accompanied by a massive database of microbial trait annotations and an interactive website for interpreting metagenomic data. As proof-of-principle, we tested TaxSEA using HMP data where it was able to accurately identify human body sites in an independent dataset with high statistical significance. Next, we demonstrated that TaxSEA was able to correctly identify disease associations using intratumoral CRC microbiome data on TCGA samples. This analysis also identified a possible microbial link between CRC and periodontal disease, which was supported by the finding that periodontitis-associated bacteria were predictive of both tumor stage and patient survival. Additionally, we found that tumor and normal tissue microbiomes have

diverging metabolic functions for processing ammonia and identified several antibiotics that may be effective in reducing dysbiotic microbiota at the tumor site.

Beyond the biomedical field, microorganisms are relevant to several other important industries, including agriculture, food processing, chemical engineering, sewage treatment, and scientific research. Like humans, plants are colonized by multitude of bacteria and fungi, each of which may play a supportive or deleterious role in plant fitness and agricultural yield [7, 375]. Additionally, recent research on fermented and cultured food production has focused on the effect of various microorganisms on food quality [376, 377]. Microbial composition can also have a significant effect on public health considerations such as water quality [378] and the efficiency of sewage treatment [379]. Beyond identifying traits that are directly relevant to healthcare and medicine, we hope that TaxSEA will be used identify traits and associated microorganisms related to agriculture, industry, and environmental sciences.

6 Conclusion

Big data and bioinformatics are becoming increasingly central to the study of biological systems [66]. This is true for nearly all realms of biology, however the increased emphasis on data science and computation is particularly relevant to the cancer microbiome field: in the modern era, the use of genome tools and sequencing technologies are all but mandatory for the study of both human cancers and human microbiomes. It has long been understood that both somatic and germline DNA mutations profoundly influence cancer; genomic analyses are thus central to cancer research efforts. Similarly, microorganisms are defined by their genomes. This is true in a very literal sense for the purposes of taxonomic classification, but also for the >99% of all species that cannot be otherwise characterized using standard culturing methods [11, 12]. Thus, sequencing the community of genomic DNA present in a given microbiome is a mandatory step in determining its composition.

As such, cancer genomics and microbiology are becoming increasingly data-driven fields. Genome sequencing methods produce terabyte-scale datasets, and even once these datasets have been processed, they frequently contain millions of datapoints. Simultaneously, the production of human sequencing data is doubling about every seven months and is expected to reach a rate of 10^{21} bases per year by 2025 [64]. This dizzying rate of data production poses significant challenges, as well as significant

opportunities. Sequenced tumors and microbial genomes are expected to comprise a large fraction of these datasets. Therefore, computational analyses that take advantage of these data scales have the potential to bring significant insights into host-microbe interactions relevant to cancer.

Tools and resources such as those presented in this work have helped to shed light on these complex dynamics. As described in Chapter 2, there are a myriad of statistical and experimental strategies for mapping out microbial interactions. While cross-sectional methods can identify associations between microorganisms, carefully designed longitudinal studies are necessary to determine the potential cause-effect relationships underlying microbiome ecosystem dynamics, a necessary step towards reproducibly modulating human microbiomes to enhance the treatment and prevention of human cancers. Furthermore, to understand the ways in which microbe-microbe interactions accumulate to affect the host, metagenomic information will need to be collected alongside multi-omic molecular profiling data, allowing summative analyses that integrate genetic, epigenetic, transcriptomic, and proteomic information. Unfortunately, such multi-omic analyses are remain prohibitively expensive for most individual research groups. Until prospective, multi-institutional analyses are planned and executed, we are left searching for alternatives.

For nearly a decade, TCGA has collected nearly 20,000 biospecimens from 33 cancer types. A significant number of these samples have been extensively analyzed, using multiple sequencing modalities as well as high-throughput proteomic profiling. These samples are accompanied by detailed clinical annotations, including tumor stage and patient survival. In Chapters 3 and 4, it is shown that sequencing data from TCGA can be used to extract bacterial and fungal metagenomic information at species-level resolution. As the microbial content of these sequencing datasets is low relative to human DNA, the microbial compositions of these samples must be carefully screened for contamination. To that end, multiple distinct strategies for identifying and mitigating contamination and false-positive signals were devised. The prevalence of bacteria between tissue and blood samples was compared, in an analysis which found that species detected at equal rates in both sample types were mostly contaminants, thus allowing isolation of the tissue-resident bacterial microbiome. Contaminant fungi, which were relatively less abundant in the tissue could not be so easily identified with this method. Instead, a horizontal and vertical analysis of genomic sequence alignments was performed, finding that horizontal distribution patterns are indicative of false-positive alignments, while vertical coverage depth across sequencing projects could be used to identify contamination. It is anticipated that such multimodal decontamination models will help to guide similar studies of low-biomass tissue samples going forward.

Once contaminant bacteria and fungi were identified and removed from these datasets, several interesting analyses were allowed. Since these bacterial and fungal profiles are matched not only to one another but also with the extensive multi-omic molecular profiling mentioned previously, some of the methods described in Chapter 2 were used to identify several interesting trans-kingdom microbial interactions, alongside informative correlations between specific microbiomes and human gene expression. Additionally, a signature of bacteria and fungi was identified in blood samples from patients with lower GI cancers which was absent in blood samples from brain cancer patients; this indicated that bacteria and fungi from lower GI tumors may translocate to the bloodstream. Several interesting associations between intratumoral microbiota and disease outcomes were also identified, together highlighting the possibility that the bacterial and fungal composition of human tumors and associated blood samples might be leveraged as a potential prognostic biomarker.

Clearly, analyses that utilize publicly available sequencing data have the potential to uncover interesting microbe-microbe and host-microbe interactions that are relevant to cancer. Beyond sequencing data, meta-analyses that leverage other forms of publicly available microbiome data can be used. In Chapter 5, several public microbiome resources were used to develop the TaxSEA database, a curated, hierarchically organized resource that contains millions of trait-microbe associations. The associated

TaxSEA analytical toolkit and website can thus be leveraged to support the interpretation microbiome profiling data. This method was validated using HMP data, and subsequently applied to differential abundance data comparing the microbiomes of tumor samples and adjacent normal tissue, obtained from TCMA database described in Chapter 3. The analysis of CRC tumor microbiomes revealed several fascinating associations, including a potential link between CRC and species implicated in periodontal disease. Such integrative analyses therefore have the potential to bring significant insights into the interactions between microbes and the host that are relevant to cancer as well as promote hypothesis generation.

Thus, each of TCMA, TCFA, and TaxSEA help to answer a distinct question, provide a novel resource, and enhance the interpretation of metagenomics data. In particular, these tools and resources have helped to shed light on the complex interactions that take place between the microbiome and human cancers. Already, the data provided by TCMA have been used in several publications exploring the relationship between the microbiome and cancer [380-385]. Beyond cancer, these works also establish broadly applicable methods for the analysis of tissue microbiomes and the interpretation metagenomic data. I hope that the tools and resources presented here will continue to remain useful to the microbiome, cancer, and cancer-microbiome research communities in the years to come.

References

1. Groussin, M., F. Mazel, and E.J. Alm, *Co-evolution and Co-speciation of Host-Gut Bacteria Systems*. *Cell Host Microbe*, 2020. **28**(1): p. 12-22.
2. Woolhouse, M.E., et al., *Biological and biomedical implications of the co-evolution of pathogens and their hosts*. *Nat Genet*, 2002. **32**(4): p. 569-77.
3. Battistuzzi, F.U., A. Feijao, and S.B. Hedges, *A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land*. *BMC Evol Biol*, 2004. **4**: p. 44.
4. Hooke, R., et al., *Micrographia : or, Some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*. 1665, London: Printed by Jo. Martyn, and Ja. Allestry, Printers to the Royal Society, and are to be sold at their Shop at the Bell, in S. Paul's Church-yard. 18 p. l., 246, 10 p.
5. Kircher, A., *Athanasii Kircheri è Soc. Jesu Scrutinium physico-medicum contagiosae luis, quae dicitur pestis : quo origo, caussae, signa, prognostica pestis nec non insolentes malignantis naturae effectus, qui statis temporibus, coelestium influxuum virtute & efficacia tum in elementis tum in epidemiis hominum animantium[ue] morbis elucescunt, una cum appropriatis remediorum antidotis nova doctrina in lucem eruuntur : cum praefatione*. 1671, Lipsiae: Sumptibus haered. Schurerianor & Joh. Fritzschi, typis Johannis Baueri. 16 , 148, 28 p.
6. Wilkinson, J.E., et al., *A framework for microbiome science in public health*. *Nat Med*, 2021. **27**(5): p. 766-774.
7. Singh, B.K., et al., *Crop microbiome and sustainable agriculture*. *Nat Rev Microbiol*, 2020. **18**(11): p. 601-602.
8. Shaw, A.J., et al., *Metabolic engineering of microbial competitive advantage for industrial fermentation processes*. *Science*, 2016. **353**(6299): p. 583-6.
9. Cohen, S.N., et al., *Construction of biologically functional bacterial plasmids in vitro*. *Proc Natl Acad Sci U S A*, 1973. **70**(11): p. 3240-4.
10. Doudna, J.A. and E. Charpentier, *Genome editing. The new frontier of genome engineering with CRISPR-Cas9*. *Science*, 2014. **346**(6213): p. 1258096.

11. Hugenholtz, P., B.M. Goebel, and N.R. Pace, *Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity*. J Bacteriol, 1998. **180**(18): p. 4765-74.
12. Thompson, L.R., et al., *A communal catalogue reveals Earth's multiscale microbial diversity*. Nature, 2017. **551**(7681): p. 457-463.
13. Huttenhower, C., et al., *Structure, function and diversity of the healthy human microbiome*. Nature, 2012. **486**(7402): p. 207-214.
14. Sender, R., S. Fuchs, and R. Milo, *Revised Estimates for the Number of Human and Bacteria Cells in the Body*. PLoS Biol, 2016. **14**(8): p. e1002533.
15. Turnbaugh, P.J., et al., *An obesity-associated gut microbiome with increased capacity for energy harvest*. Nature, 2006. **444**(7122): p. 1027-31.
16. Giongo, A., et al., *Toward defining the autoimmune microbiome for type 1 diabetes*. ISME J, 2011. **5**(1): p. 82-91.
17. Gevers, D., et al., *The treatment-naïve microbiome in new-onset Crohn's disease*. Cell Host Microbe, 2014. **15**(3): p. 382-392.
18. Morgan, X.C., et al., *Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment*. Genome Biol, 2012. **13**(9): p. R79.
19. Foster, J.A. and K.A. McVey Neufeld, *Gut-brain axis: how the microbiome influences anxiety and depression*. Trends Neurosci, 2013. **36**(5): p. 305-12.
20. Zheng, P., et al., *The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice*. Sci Adv, 2019. **5**(2): p. eaau8317.
21. Vuong, H.E. and E.Y. Hsiao, *Emerging Roles for the Gut Microbiome in Autism Spectrum Disorder*. Biol Psychiatry, 2017. **81**(5): p. 411-423.
22. Sampson, T.R., et al., *Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease*. Cell, 2016. **167**(6): p. 1469-1480 e12.
23. Ebbell, B., *The Papyrus Ebers, the greatest Egyptian medical document*. 1937, Copenhagen,

London,; Levin & Munksgaard;
H. Milford, Oxford university press. 135 p.

24. Starnes, C.O., *Coley's toxins in perspective*. Nature, 1992. **357**(6373): p. 11-2.
25. Sepich-Poore, G.D., et al., *The microbiome and human cancer*. Science, 2021. **371**(6536).
26. *Livingston-Wheeler therapy*. CA Cancer J Clin, 1990. **40**(2): p. 103-8.
27. Rous, P., *A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells*. J Exp Med, 1911. **13**(4): p. 397-411.
28. Bullman, S., et al., *Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer*. Science, 2017. **358**(6369): p. 1443-1448.
29. Pleguezuelos-Manzano, C., et al., *Mutational signature in colorectal cancer caused by genotoxic pks(+) E. coli*. Nature, 2020. **580**(7802): p. 269-273.
30. Matson, V., et al., *The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients*. Science, 2018. **359**(6371): p. 104-108.
31. Riquelme, E., et al., *Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes*. Cell, 2019. **178**(4): p. 795-806 e12.
32. Viaud, S., et al., *The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide*. Science, 2013. **342**(6161): p. 971-6.
33. Sivan, A., et al., *Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy*. Science, 2015. **350**(6264): p. 1084-9.
34. Gopalakrishnan, V., et al., *Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients*. Science, 2018. **359**(6371): p. 97-103.
35. Nejman, D., et al., *The human tumor microbiome is composed of tumor type-specific intracellular bacteria*. Science, 2020. **368**(6494): p. 973-980.
36. Dohlman, A.B., et al., *The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants*. Cell Host Microbe, 2021. **29**(2): p. 281-298 e5.

37. Aykut, B., et al., *The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL*. *Nature*, 2019. **574**(7777): p. 264-267.
38. Jin, C., et al., *Commensal Microbiota Promote Lung Cancer Development via gammadelta T Cells*. *Cell*, 2019. **176**(5): p. 998-1013 e16.
39. Kostic, A.D., et al., *Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment*. *Cell Host Microbe*, 2013. **14**(2): p. 207-15.
40. Humans, I.W.G.o.t.E.o.C.R.t., *Biological agents. Volume 100 B. A review of human carcinogens*. IARC Monogr Eval Carcinog Risks Hum, 2012. **100**(Pt B): p. 1-441.
41. Polk, D.B. and R.M. Peek, Jr., *Helicobacter pylori: gastric cancer and beyond*. *Nat Rev Cancer*, 2010. **10**(6): p. 403-14.
42. Boleij, A., et al., *The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients*. *Clin Infect Dis*, 2015. **60**(2): p. 208-15.
43. Kostic, A.D., et al., *Genomic analysis identifies association of Fusobacterium with colorectal carcinoma*. *Genome Res*, 2012. **22**(2): p. 292-8.
44. Castellarin, M., et al., *Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma*. *Genome Res*, 2012. **22**(2): p. 299-306.
45. Schmidt, B.L., et al., *Changes in abundance of oral microbiota associated with oral cancer*. *PLoS One*, 2014. **9**(6): p. e98741.
46. Peters, B.A., et al., *Oral Microbiome Composition Reflects Prospective Risk for Esophageal Cancers*. *Cancer Res*, 2017. **77**(23): p. 6777-6787.
47. Yamamura, K., et al., *Human Microbiome Fusobacterium Nucleatum in Esophageal Cancer Tissue Is Associated with Prognosis*. *Clin Cancer Res*, 2016. **22**(22): p. 5574-5581.
48. Hieken, T.J., et al., *The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease*. *Sci Rep*, 2016. **6**: p. 30751.
49. Greathouse, K.L., et al., *Interaction between the microbiome and TP53 in human lung cancer*. *Genome Biol*, 2018. **19**(1): p. 123.

50. Nene, N.R., et al., *Association between the cervicovaginal microbiome, BRCA1 mutation status, and risk of ovarian cancer: a case-control study*. *Lancet Oncol*, 2019. **20**(8): p. 1171-1182.
51. Honda, K. and D.R. Littman, *The microbiota in adaptive immune homeostasis and disease*. *Nature*, 2016. **535**(7610): p. 75-84.
52. Zitvogel, L., et al., *The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies*. *Science*, 2018. **359**(6382): p. 1366-1370.
53. Vetizou, M., et al., *Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota*. *Science*, 2015. **350**(6264): p. 1079-84.
54. Vergara, D., et al., *The Cancer Microbiota: EMT and Inflammation as Shared Molecular Mechanisms Associated with Plasticity and Progression*. *J Oncol*, 2019. **2019**: p. 1253727.
55. Hofman, P. and V. Vouret-Craviari, *Microbes-induced EMT at the crossroad of inflammation and cancer*. *Gut Microbes*, 2012. **3**(3): p. 176-85.
56. Martin, T.A. and W.G. Jiang, *Loss of tight junction barrier function and its role in cancer metastasis*. *Biochim Biophys Acta*, 2009. **1788**(4): p. 872-91.
57. Squarzanti, D.F., et al., *Non-Melanoma Skin Cancer: news from microbiota research*. *Crit Rev Microbiol*, 2020. **46**(4): p. 433-449.
58. Steck, S.E. and E.A. Murphy, *Dietary patterns and cancer risk*. *Nat Rev Cancer*, 2020. **20**(2): p. 125-138.
59. Wirbel, J., et al., *Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer*. *Nat Med*, 2019. **25**(4): p. 679-689.
60. Klein, R.S., et al., *Association of *Streptococcus bovis* with carcinoma of the colon*. *N Engl J Med*, 1977. **297**(15): p. 800-2.
61. Garrett, W.S., *Cancer and the microbiota*. *Science*, 2015. **348**(6230): p. 80-6.
62. Poore, G.D., et al., *Microbiome analyses of blood and tissues suggest cancer diagnostic approach*. *Nature*, 2020. **579**(7800): p. 567-574.

63. Hanahan, D., *Hallmarks of Cancer: New Dimensions*. *Cancer Discov*, 2022. **12**(1): p. 31-46.
64. Stephens, Z.D., et al., *Big Data: Astronomical or Genomical?* *PLoS Biol*, 2015. **13**(7): p. e1002195.
65. Russell, P.H., et al., *A large-scale analysis of bioinformatics code on GitHub*. *PLoS One*, 2018. **13**(10): p. e0205898.
66. Chen, X.W. and J.X. Gao, *Big Data Bioinformatics*. *Methods*, 2016. **111**: p. 1-2.
67. Martiny, J.B., et al., *Microbiomes in light of traits: A phylogenetic perspective*. *Science*, 2015. **350**(6261): p. aac9323.
68. Hall, E.K., et al., *Understanding how microbiomes influence the systems they inhabit*. *Nat Microbiol*, 2018. **3**(9): p. 977-982.
69. Janssens, Y., et al., *Disbiome database: linking the microbiome to disease*. *BMC Microbiol*, 2018. **18**(1): p. 50.
70. Kou, Y., et al., *Microbe-set enrichment analysis facilitates functional interpretation of microbiome profiling data*. *Sci Rep*, 2020. **10**(1): p. 21466.
71. Kostic, A.D., et al., *PathSeq: software to identify or discover microbes by deep sequencing of human tissue*. *Nat Biotechnol*, 2011. **29**(5): p. 393-6.
72. Robinson, K.M., et al., *Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data*. *Microbiome*, 2017. **5**(1): p. 9.
73. Lu, J. and S.L. Salzberg, *Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2*. *Microbiome*, 2020. **8**(1): p. 124.
74. Eisenhofer, R., et al., *Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations*. *Trends Microbiol*, 2019. **27**(2): p. 105-117.
75. Davis, N.M., et al., *Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data*. *Microbiome*, 2018. **6**(1): p. 226.

76. Glassing, A., et al., *Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples*. *Gut Pathog*, 2016. **8**: p. 24.
77. Salter, S.J., et al., *Reagent and laboratory contamination can critically impact sequence-based microbiome analyses*. *BMC Biol*, 2014. **12**: p. 87.
78. Baquero, F. and C. Nombela, *The microbiome as a human organ*. *Clinical Microbiology and Infection*, 2012. **18**: p. 2-4.
79. O'Hara, A.M. and F. Shanahan, *The gut flora as a forgotten organ*. *Embo Reports*, 2006. **7**(7): p. 688-693.
80. Hattori, M. and T.D. Taylor, *The Human Intestinal Microbiome: A New Frontier of Human Biology*. *DNA Research*, 2009. **16**(1): p. 1-12.
81. Perry, R.J., et al., *Acetate mediates a microbiome-brain-beta-cell axis to promote metabolic syndrome*. *Nature*, 2016. **534**(7606): p. 213-+.
82. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins*. *Nature*, 2009. **457**(7228): p. 480-U7.
83. Pop, M., et al., *Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition*. *Genome Biology*, 2014. **15**(6).
84. Gulden, E., F.S. Wong, and L. Wen, *The gut microbiota and Type 1 Diabetes*. *Clinical Immunology*, 2015. **159**(2): p. 143-153.
85. Jalanka-Tuovinen, J., et al., *Faecal microbiota composition and host-microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome*. *Gut*, 2014. **63**(11): p. 1737-1745.
86. Morgan, X.C., et al., *Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment*. *Genome Biology*, 2012. **13**(9).
87. Pascal, V., et al., *A microbial signature for Crohn's disease*. *Gut*, 2017. **66**(5): p. 813-822.
88. Yu, Y.N. and J.Y. Fang, *Gut Microbiota and Colorectal Cancer*. *Gastrointestinal Tumors*, 2015. **2**(1): p. 26-32.

89. Mulle, J.G., W.G. Sharp, and J.F. Cubells, *The Gut Microbiome: A New Frontier in Autism Research*. Current Psychiatry Reports, 2013. **15**(2).
90. Li, Q.R., et al., *The Gut Microbiota and Autism Spectrum Disorders*. Frontiers in Cellular Neuroscience, 2017. **11**.
91. Dickerson, F., E. Severance, and R. Yolken, *The microbiome, immunity, and schizophrenia and bipolar disorder*. Brain Behavior and Immunity, 2017. **62**: p. 46-52.
92. Sampson, T.R., et al., *Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease*. Cell, 2016. **167**(6): p. 1469-+.
93. Winter, G., et al., *Gut microbiome and depression: what we know and what we need to know*. Reviews in the Neurosciences, 2018. **29**(6): p. 629-643.
94. Xu, M.Q., et al., *Fecal microbiota transplantation broadening its application beyond intestinal disorders*. World Journal of Gastroenterology, 2015. **21**(1): p. 102-111.
95. Kang, D.W., et al., *Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study*. Microbiome, 2017. **5**.
96. Evrensel, A. and M.E. Ceylan, *Fecal Microbiota Transplantation and Its Usage in Neuropsychiatric Disorders*. Clinical Psychopharmacology and Neuroscience, 2016. **14**(3): p. 231-237.
97. Allen, S.J., et al., *Probiotics in the prevention of eczema: a randomised controlled trial*. Archives of Disease in Childhood, 2014. **99**(11): p. 1014-1019.
98. Allen, S.J., et al., *Lactobacilli and bifidobacteria in the prevention of antibiotic-associated diarrhoea and Clostridium difficile diarrhoea in older inpatients (PLACIDE): a randomised, double-blind, placebo-controlled, multicentre trial*. Lancet, 2013. **382**(9900): p. 1249-1257.
99. Schnadower, D., et al., *Randomised controlled trial of Lactobacillus rhamnosus (LGG) versus placebo in children presenting to the emergency department with acute gastroenteritis: the PECARN probiotic study protocol*. Bmj Open, 2017. **7**(9).
100. Freedman, S.B., et al., *Impact of emergency department probiotic treatment of pediatric gastroenteritis: study protocol for the PROGUT (Probiotic Regimen for Outpatient*

- Gastroenteritis Utility of Treatment*) randomized controlled trial. *Trials*, 2014. **15**: p. 170.
101. De Smet, R. and K. Marchal, *Advantages and limitations of current network inference methods*. *Nature Reviews Microbiology*, 2010. **8**(10): p. 717-729.
 102. Veiga, D.F.T., B. Dutta, and G. Balazsi, *Network inference and network response identification: moving genome-scale data to the next level of biological discovery*. *Molecular Biosystems*, 2010. **6**(3): p. 469-480.
 103. Lovell, D., et al., *Proportions, percentages, ppm: do the molecular biosciences treat compositional data right?* *Compositional Data Analysis: Theory and Applications*, 2011: p. 193-207.
 104. Tsilimigras, M.C.B. and A.A. Fodor, *Compositional data analysis of the microbiome: fundamentals, tools, and challenges*. *Annals of Epidemiology*, 2016. **26**(5): p. 330-335.
 105. Aitchison, J., *The Statistical-Analysis of Compositional Data*. *Journal of the Royal Statistical Society Series B-Methodological*, 1982. **44**(2): p. 139-177.
 106. Li, H.Z., *Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis*. *Annual Review of Statistics and Its Application*, Vol 2, 2015. **2**: p. 73-94.
 107. Friedman, J. and E.J. Alm, *Inferring Correlation Networks from Genomic Survey Data*. *Plos Computational Biology*, 2012. **8**(9).
 108. Ilhan, Z.E., et al., *pH-Mediated Microbial and Metabolic Interactions in Fecal Enrichment Cultures*. *mSphere*, 2017. **2**(3).
 109. Faust, K. and J. Raes, *Microbial interactions: from networks to models*. *Nature Reviews Microbiology*, 2012. **10**(8): p. 538-550.
 110. Faust, K. and J. Raes, *Microbial interactions: from networks to models*. *Nat Rev Microbiol*, 2012. **10**(8): p. 538-50.
 111. Xiao, Y.D., et al., *Mapping the ecological networks of microbial communities*. *Nature Communications*, 2017. **8**.
 112. Weiss, S., et al., *Correlation detection strategies in microbial data sets vary widely in sensitivity and precision*. *ISME J*, 2016. **10**(7): p. 1669-81.

113. Pawlowsky-Glahn, V., J.J. Egozcue, and R. Tolosana-Delgado, *Modeling and Analysis of Compositional Data Introduction*. Modeling and Analysis of Compositional Data, 2015: p. 1-7.
114. Kurtz, Z.D., et al., *Sparse and compositionally robust inference of microbial ecological networks*. PLoS Comput Biol, 2015. **11**(5): p. e1004226.
115. Tipton, L., et al., *Fungi stabilize connectivity in the lung and skin microbial ecosystems*. Microbiome, 2018. **6**(1): p. 12.
116. Fang, H., et al., *CCLasso: correlation inference for compositional data through Lasso*. Bioinformatics, 2015. **31**(19): p. 3172-80.
117. Ban, Y., L. An, and H. Jiang, *Investigating microbial co-occurrence patterns based on metagenomic compositional data*. Bioinformatics, 2015. **31**(20): p. 3322-9.
118. Schwager, E., et al., *A Bayesian method for detecting pairwise associations in compositional data*. PLoS Comput Biol, 2017. **13**(11): p. e1005852.
119. Lo, C. and R. Marculescu, *MPLasso: Inferring microbial association networks using prior microbial knowledge*. PLoS Comput Biol, 2017. **13**(12): p. e1005915.
120. Reshef, D.N., et al., *Detecting novel associations in large data sets*. Science, 2011. **334**(6062): p. 1518-24.
121. Ruan, Q., et al., *Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors*. Bioinformatics, 2006. **22**(20): p. 2532-8.
122. Ki, B.M., H.W. Ryu, and K.S. Cho, *Extended local similarity analysis (eLSA) reveals unique associations between bacterial community structure and odor emission during pig carcasses decomposition*. J Environ Sci Health A Tox Hazard Subst Environ Eng, 2018. **53**(8): p. 718-727.
123. Xia, L.C., et al., *Efficient statistical significance approximation for local similarity analysis of high-throughput time series data*. Bioinformatics, 2013. **29**(2): p. 230-7.
124. Deng, Y., et al., *Molecular ecological network analyses*. BMC Bioinformatics, 2012. **13**: p. 113.

125. Faust, K., et al., *Microbial co-occurrence relationships in the human microbiome*. PLoS Comput Biol, 2012. **8**(7): p. e1002606.
126. Cusco, A., et al., *Individual signatures and environmental factors shape skin microbiota in healthy dogs*. Microbiome, 2017. **5**(1): p. 139.
127. Potgens, S.A., et al., *Klebsiella oxytoca expands in cancer cachexia and acts as a gut pathobiont contributing to intestinal dysfunction*. Sci Rep, 2018. **8**(1): p. 12321.
128. Caporaso, J.G., et al., *Moving pictures of the human microbiome*. Genome Biol, 2011. **12**(5): p. R50.
129. Gourevitch, B., R.L. Bouquin-Jeannes, and G. Faucon, *Linear and nonlinear causality between signals: methods, examples and neurophysiological applications*. Biol Cybern, 2006. **95**(4): p. 349-69.
130. Gibbons, S.M., et al., *Two dynamic regimes in the human gut microbiome*. PLoS Comput Biol, 2017. **13**(2): p. e1005364.
131. Sugihara, G., et al., *Detecting causality in complex ecosystems*. Science, 2012. **338**(6106): p. 496-500.
132. Granger, C.W.J., *Investigating Causal Relations by Econometric Models and Cross-Spectral Methods*. Econometrica, 1969. **37**(3): p. 424-438.
133. Eichler, M., *Causal inference with multiple time series: principles and problems*. Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences, 2013. **371**(1997).
134. Baksi, K.D., B.K. Kuntal, and S.S. Mande, *'TIME': A Web Application for Obtaining Insights into Microbial Ecology Using Longitudinal Microbiome Data*. Frontiers in Microbiology, 2018. **9**.
135. Lotka, A.J., *Contribution to the theory of periodic reactions*. Journal of Physical Chemistry, 1910. **14**(3): p. 271-274.
136. Volterra, V., *Fluctuations in the abundance of a species considered mathematically*. Nature, 1926. **118**: p. 558-560.

137. Goerges, S., et al., *Commercial ripening starter microorganisms inoculated into cheese milk do not successfully establish themselves in the resident microbial ripening consortia of a South German red smear cheese*. *Applied and Environmental Microbiology*, 2008. **74**(7): p. 2210-2217.
138. Stein, R.R., et al., *Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota*. *Plos Computational Biology*, 2013. **9**(12).
139. Fisher, C.K. and P. Mehta, *Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression*. *PLoS One*, 2014. **9**(7): p. e102451.
140. Shaw, G.T., Y.Y. Pao, and D. Wang, *MetaMIS: a metagenomic microbial interaction simulator based on microbial community profiles*. *BMC Bioinformatics*, 2016. **17**(1): p. 488.
141. Gonze, D., et al., *Microbial communities as dynamical systems*. *Current Opinion in Microbiology*, 2018. **44**: p. 41-49.
142. Moree, W.J., et al., *Interkingdom metabolic transformations captured by microbial imaging mass spectrometry*. *Proceedings of the National Academy of Sciences of the United States of America*, 2012. **109**(34): p. 13811-13816.
143. Momeni, B., L. Xie, and W.Y. Shou, *Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions*. *Elife*, 2017. **6**.
144. Bar-Joseph, Z., A. Gitter, and I. Simon, *Studying and modelling dynamic biological processes using time-series gene expression data*. *Nat Rev Genet*, 2012. **13**(8): p. 552-64.
145. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo*. *Genome Biol*, 2006. **7**(5): p. R36.
146. Aijo, T., et al., *Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation*. *Bioinformatics*, 2014. **30**(12): p. i113-20.
147. Gerber, G.K., A.B. Onderdonk, and L. Bry, *Inferring dynamic signatures of microbes in complex host ecosystems*. *PLoS Comput Biol*, 2012. **8**(8): p. e1002624.

148. Bucci, V., et al., *MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses*. *Genome Biol*, 2016. **17**(1): p. 121.
149. Aijo, T., C.L. Muller, and R. Bonneau, *Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing*. *Bioinformatics*, 2018. **34**(3): p. 372-380.
150. Chen, J. and H. Li, *Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis*. *Ann Appl Stat*, 2013. **7**(1).
151. West, M. and J. Harrison, *Bayesian forecasting and dynamic models*. 2nd ed. Springer series in statistics. 1997, New York: Springer. xiv, 680 p.
152. Stein, R.R., et al., *Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota*. *PLoS Comput Biol*, 2013. **9**(12): p. e1003388.
153. Marino, S., et al., *Mathematical modeling of primary succession of murine intestinal microbiota*. *Proc Natl Acad Sci U S A*, 2014. **111**(1): p. 439-44.
154. Buffie, C.G., et al., *Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile**. *Nature*, 2015. **517**(7533): p. 205-8.
155. Friedman, J. and E.J. Alm, *Inferring correlation networks from genomic survey data*. *PLoS Comput Biol*, 2012. **8**(9): p. e1002687.
156. Bucci, V. and J.B. Xavier, *Towards predictive models of the human gut microbiome*. *J Mol Biol*, 2014. **426**(23): p. 3907-16.
157. Gerber, G.K., *The dynamic microbiome*. *FEBS Lett*, 2014. **588**(22): p. 4131-9.
158. Datta, M.S., et al., *Microbial interactions lead to rapid micro-scale successions on model marine particles*. *Nat Commun*, 2016. **7**: p. 11965.
159. Angulo, M.T., et al., *Fundamental limitations of network reconstruction from temporal data*. *Journal of the Royal Society Interface*, 2017. **14**(127).
160. Liang, X., F.D. Bushman, and G.A. FitzGerald, *Rhythmicity of the intestinal microbiota is regulated by gender and the host circadian clock*. *Proc Natl Acad Sci U S A*, 2015. **112**(33): p. 10479-84.

161. Trosvik, P., et al., *Characterizing mixed microbial population dynamics using time-series analysis*. ISME J, 2008. **2**(7): p. 707-15.
162. Trosvik, P., et al., *Web of ecological interactions in an experimental gut microbiota*. Environ Microbiol, 2010. **12**(10): p. 2677-87.
163. Blaut, M., *Ecology and physiology of the intestinal tract*. Curr Top Microbiol Immunol, 2013. **358**: p. 247-72.
164. Mackie, R.I., B.A. White, and R.E. Isaacson, *Gastrointestinal microbiology*. Chapman & Hall microbiology series. 1997, New York: Chapman & Hall.
165. Phelan, V.V., et al., *Microbial metabolic exchange-the chemotype-to-phenotype link*. Nature Chemical Biology, 2012. **8**(1): p. 26-35.
166. Rosenfeld, C.S., *Gut Dysbiosis in Animals Due to Environmental Chemical Exposures*. Frontiers in Cellular and Infection Microbiology, 2017. **7**.
167. Chong, J. and J. Xia, *Computational Approaches for Integrative Analysis of the Metabolome and Microbiome*. Metabolites, 2017. **7**(4).
168. Dhariwal, A., et al., *MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data*. Nucleic Acids Research, 2017. **45**(W1): p. W180-W188.
169. Larsen, P.E., et al., *PREDICTED RELATIVE METABOLOMIC TURNOVER Predicting Changes in the Environmental Metabolome from the Metagenome*. Bioinformatics 2011, 2011: p. 337-345.
170. Abubucker, S., et al., *Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome*. Plos Computational Biology, 2012. **8**(6).
171. Langille, M.G.I., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences*. Nature Biotechnology, 2013. **31**(9): p. 814-+.
172. Asshauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data*. Bioinformatics, 2015. **31**(17): p. 2882-2884.
173. Iwai, S., et al., *Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes*. PLoS One, 2016. **11**(11): p. e0166104.

174. Franzosa, E.A., et al., *Relating the metatranscriptome and metagenome of the human gut*. Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(22): p. E2329-E2338.
175. Verberkmoes, N.C., et al., *Shotgun metaproteomics of the human distal gut microbiota*. Isme Journal, 2009. **3**(2): p. 179-189.
176. Perez-Cobas, A.E., et al., *Gut microbiota disturbance during antibiotic therapy: a multi-omic approach*. Gut, 2013. **62**(11): p. 1591-601.
177. Noecker, C., et al., *Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation*. Msystems, 2016. **1**(1).
178. Trygg, J. and S. Wold, *O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter*. Journal of Chemometrics, 2003. **17**(1): p. 53-64.
179. Hotelling, H., *Relations between two sets of variates*. Biometrika, 1936. **28**: p. 321-377.
180. Doledec, S. and D. Chessel, *Co-Inertia Analysis - an Alternative Method for Studying Species Environment Relationships*. Freshwater Biology, 1994. **31**(3): p. 277-294.
181. Chong, J. and J.G. Xia, *Computational Approaches for Integrative Analysis of the Metabolome and Microbiome*. Metabolites, 2017. **7**(4).
182. Sung, J., et al., *Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis*. Nat Commun, 2017. **8**: p. 15393.
183. Donohoe, D.R., et al., *The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon*. Cell Metab, 2011. **13**(5): p. 517-26.
184. Furusawa, Y., et al., *Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells*. Nature, 2013. **504**(7480): p. 446-50.
185. Louis, P., G.L. Hold, and H.J. Flint, *The gut microbiota, bacterial metabolites and colorectal cancer*. Nat Rev Microbiol, 2014. **12**(10): p. 661-72.
186. Thiele, I., et al., *A community-driven global reconstruction of human metabolism*. Nat Biotechnol, 2013. **31**(5): p. 419-25.

187. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction*. Nat Protoc, 2010. **5**(1): p. 93-121.
188. Oberhardt, M.A., B.O. Palsson, and J.A. Papin, *Applications of genome-scale metabolic reconstructions*. Mol Syst Biol, 2009. **5**: p. 320.
189. Knight, J.M., et al., *Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing*. Scientific Reports, 2014. **4**.
190. Shah, P., et al., *A microfluidics-based in vitro model of the gastrointestinal human-microbe interface*. Nature Communications, 2016. **7**.
191. Kostic, A.D., M.R. Howitt, and W.S. Garrett, *Exploring host-microbiota interactions in animal models and humans*. Genes & Development, 2013. **27**(7): p. 701-718.
192. Yatsunenkov, T., et al., *Human gut microbiome viewed across age and geography*. Nature, 2012. **486**(7402): p. 222-+.
193. Gaulke, C.A. and T.J. Sharpton, *The influence of ethnicity and geography on human gut microbiome composition*. Nature Medicine, 2018. **24**(10): p. 1495-1496.
194. David, L.A., et al., *Diet rapidly and reproducibly alters the human gut microbiome*. Nature, 2014. **505**(7484): p. 559-+.
195. Tung, J., et al., *Social networks predict gut microbiome composition in wild baboons*. Elife, 2015. **4**.
196. Zhang, C.H., et al., *Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice (vol 4, pg 232, 2010)*. Isme Journal, 2010. **4**(2): p. 312-313.
197. Wang, X.H., et al., *Multivariate Approach for Studying Interactions between Environmental Variables and Microbial Communities*. Plos One, 2012. **7**(11).
198. Luckey, T.D., *Introduction to intestinal microecology*. Am J Clin Nutr, 1972. **25**(12): p. 1292-4.
199. Human Microbiome Project, C., *Structure, function and diversity of the healthy human microbiome*. Nature, 2012. **486**(7402): p. 207-14.

200. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing*. Nature, 2010. **464**(7285): p. 59-65.
201. Levy, M., et al., *Dysbiosis and the immune system*. Nat Rev Immunol, 2017. **17**(4): p. 219-232.
202. Elinav, E., et al., *The cancer microbiome*. Nat Rev Cancer, 2019. **19**(7): p. 371-376.
203. Iliev, I.D. and I. Leonardi, *Fungal dysbiosis: immunity and interactions at mucosal barriers*. Nature Reviews Immunology, 2017. **17**(10): p. 635-646.
204. Schirmer, M., et al., *Microbial genes and pathways in inflammatory bowel disease*. Nat Rev Microbiol, 2019. **17**(8): p. 497-511.
205. Prast-Nielsen, S., et al., *Investigation of the skin microbiome: swabs vs. biopsies*. Br J Dermatol, 2019. **181**(3): p. 572-579.
206. Grice, E.A., et al., *A diversity profile of the human skin microbiota*. Genome Res, 2008. **18**(7): p. 1043-50.
207. Flanagan, L., et al., *Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome*. Eur J Clin Microbiol Infect Dis, 2014. **33**(8): p. 1381-90.
208. Yu, T., et al., *Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy*. Cell, 2017. **170**(3): p. 548-563 e16.
209. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of gastric adenocarcinoma*. Nature, 2014. **513**(7517): p. 202-9.
210. Cancer Genome Atlas Research, N., et al., *Integrated genomic and molecular characterization of cervical cancer*. Nature, 2017. **543**(7645): p. 378-384.
211. Tang, K.W., et al., *The landscape of viral expression and host gene fusion and adaptation in human cancer*. Nat Commun, 2013. **4**: p. 2513.
212. Choi, J.H., S.E. Hong, and H.G. Woo, *Pan-cancer analysis of systematic batch effects on somatic sequence variations*. BMC Bioinformatics, 2017. **18**(1): p. 211.

213. CDC. *NHSN Organism List*. National Healthcare Safety Network (NHSN) Patient Safety Component Manual 2019 [cited 2019 May 10]; Available from: https://www.cdc.gov/nhsn/pdfs/pscmanual/pcsmanual_current.pdf.
214. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2*. *Genome Biol*, 2019. **20**(1): p. 257.
215. Chelakkot, C., J. Ghim, and S.H. Ryu, *Mechanisms regulating intestinal barrier integrity and its pathological implications*. *Exp Mol Med*, 2018. **50**(8): p. 103.
216. Sriswasdi, S., C.C. Yang, and W. Iwasaki, *Generalist species drive microbial dispersion and evolution*. *Nat Commun*, 2017. **8**(1): p. 1162.
217. Yamamoto, Y., et al., *The *Escherichia coli* ldcC gene encodes another lysine decarboxylase, probably a constitutive enzyme*. *Genes & Genetic Systems*, 1997. **72**(3): p. 167-172.
218. Kikuchi, Y., et al., *Characterization of a second lysine decarboxylase isolated from Escherichia coli*. *Journal of bacteriology*, 1997. **179**(14): p. 4486-4492.
219. Lemonnier, M. and D. Lane, *Expression of the second lysine decarboxylase gene of Escherichia coli*. *Microbiology*, 1998. **144**(3): p. 751-760.
220. Arthur, J.C., et al., *Intestinal inflammation targets cancer-inducing activity of the microbiota*. *Science*, 2012. **338**(6103): p. 120-3.
221. Koskiniemi, S., et al., *Selection-driven gene loss in bacteria*. *PLoS Genet*, 2012. **8**(6): p. e1002787.
222. Mira, A., H. Ochman, and N.A. Moran, *Deletional bias and the evolution of bacterial genomes*. *Trends Genet*, 2001. **17**(10): p. 589-96.
223. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. *Nat Protoc*, 2009. **4**(1): p. 44-57.
224. Porcheron, G., et al., *Iron, copper, zinc, and manganese transport and regulation in pathogenic Enterobacteria: correlations between strains, site of infection and the relative importance of the different metal transport systems for virulence*. *Front Cell Infect Microbiol*, 2013. **3**: p. 90.

225. Rensing, C. and G. Grass, *Escherichia coli* mechanisms of copper homeostasis in a changing environment. *FEMS Microbiol Rev*, 2003. **27**(2-3): p. 197-213.
226. Yang, J.H., et al., *Widespread inosine-containing mRNA in lymphocytes regulated by ADAR1 in response to inflammation*. *Immunology*, 2003. **109**(1): p. 15-23.
227. Chung, H., et al., *Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown*. *Cell*, 2018. **172**(4): p. 811-824 e14.
228. Nossa, C.W., et al., *Activation of the abundant nuclear factor poly(ADP-ribose) polymerase-1 by Helicobacter pylori*. *Proc Natl Acad Sci U S A*, 2009. **106**(47): p. 19998-20003.
229. Purcell, R.V., et al., *Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer*. *Sci Rep*, 2017. **7**(1): p. 11590.
230. Warren, R.L., et al., *Co-occurrence of anaerobic bacteria in colorectal carcinomas*. *Microbiome*, 2013. **1**(1): p. 16.
231. O'Donovan, D., et al., *Campylobacter ureolyticus: a portrait of the pathogen*. *Virulence*, 2014. **5**(4): p. 498-506.
232. Bullman, S., et al., *Genomic investigation into strain heterogeneity and pathogenic potential of the emerging gastrointestinal pathogen Campylobacter ureolyticus*. *PLoS One*, 2013. **8**(8): p. e71515.
233. Liu, J., et al., *An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics*. *Cell*, 2018. **173**(2): p. 400-416 e11.
234. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. *Proc Natl Acad Sci U S A*, 2005. **102**(43): p. 15545-50.
235. Abdulmir, A.S., R.R. Hafidh, and F. Abu Bakar, *The association of Streptococcus bovis/gallolyticus with colorectal tumors: the nature and the underlying mechanisms of its etiological role*. *J Exp Clin Cancer Res*, 2011. **30**: p. 11.
236. Oshima, T. and H. Miwa, *Gastrointestinal mucosal barrier function and diseases*. *Journal of Gastroenterology*, 2016. **51**(8): p. 768-778.

237. Yu, L.C., *Microbiota dysbiosis and barrier dysfunction in inflammatory bowel disease and colorectal cancers: exploring a common ground hypothesis*. J Biomed Sci, 2018. **25**(1): p. 79.
238. Brighenti, E., et al., *Interleukin 6 downregulates p53 expression and activity by stimulating ribosome biogenesis: a new pathway connecting inflammation to cancer*. Oncogene, 2014. **33**(35): p. 4396-406.
239. Gemmell, E. and G.J. Seymour, *Interleukin 1, interleukin 6 and transforming growth factor-beta production by human gingival mononuclear cells following stimulation with Porphyromonas gingivalis and Fusobacterium nucleatum*. J Periodontal Res, 1993. **28**(2): p. 122-9.
240. Rubinstein, M.R., et al., *Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin*. Cell Host Microbe, 2013. **14**(2): p. 195-206.
241. Islami, F. and F. Kamangar, *Helicobacter pylori and esophageal cancer risk: a meta-analysis*. Cancer Prev Res (Phila), 2008. **1**(5): p. 329-38.
242. Plottel, C.S. and M.J. Blaser, *Microbiome and malignancy*. Cell Host Microbe, 2011. **10**(4): p. 324-35.
243. Pertea, G. and M. Pertea, *GFF Utilities: GffRead and GffCompare [version 1; peer review: awaiting peer review]*. F1000Research, 2020. **9**(304).
244. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics, 2013. **29**(1): p. 15-21.
245. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features*. Bioinformatics, 2014. **30**(7): p. 923-30.
246. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
247. Ramírez, F., et al., *deepTools: a flexible platform for exploring deep-sequencing data*. Nucleic Acids Res, 2014. **42**(Web Server issue): p. W187-91.

248. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. *Bioinformatics*, 2010. **26**(6): p. 841-2.
249. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics*. *Genome Res*, 2009. **19**(9): p. 1639-45.
250. Bush, S.J., et al., *Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines*. *Gigascience*, 2020. **9**(2).
251. Marotz, C., et al., *DNA extraction for streamlined metagenomics of diverse environmental samples*. *Biotechniques*, 2017. **62**(6): p. 290-293.
252. Jiang, W., *A protocol for quantizing total bacterial 16S rDNA in plasma as a marker of microbial translocation in vivo*. *Cell Mol Immunol*, 2018. **15**(10): p. 937-939.
253. Bolyen, E., et al., *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2*. *Nat Biotechnol*, 2019. **37**(8): p. 852-857.
254. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB*. *Appl Environ Microbiol*, 2006. **72**(7): p. 5069-72.
255. Ma, W., et al., *MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes*. *Sci Rep*, 2017. **7**: p. 40200.
256. Foster, Z.S., T.J. Sharpton, and N.J. Grunwald, *Metacoder: An R package for visualization and manipulation of community taxonomic diversity data*. *PLoS Comput Biol*, 2017. **13**(2): p. e1005404.
257. Davidson-Pilon, C., et al., *CamDavidsonPilon/lifelines: 0.24.6*. 2020, Zenodo.
258. Davar, D., et al., *Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients*. *Science*, 2021. **371**(6529): p. 595-602.
259. Dzutsev, A., et al., *Microbes and Cancer*. *Annu Rev Immunol*, 2017. **35**: p. 199-228.
260. Finlay, B.B., et al., *Can we harness the microbiota to enhance the efficacy of cancer immunotherapy?* *Nat Rev Immunol*, 2020. **20**(9): p. 522-528.
261. Garrett, W.S., *The gut microbiota and colon cancer*. *Science*, 2019. **364**(6446): p. 1133-1135.

262. Grivennikov, S.I., F.R. Greten, and M. Karin, *Immunity, inflammation, and cancer*. Cell, 2010. **140**(6): p. 883-99.
263. Iida, N., et al., *Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment*. Science, 2013. **342**(6161): p. 967-70.
264. Routy, B., et al., *Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors*. Science, 2018. **359**(6371): p. 91-97.
265. Sharma, P., et al., *Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy*. Cell, 2017. **168**(4): p. 707-723.
266. Shiao, S.L., et al., *Commensal bacteria and fungi differentially regulate tumor responses to radiation therapy*. Cancer Cell, 2021. **39**(9): p. 1202-1213 e6.
267. Spencer, C.N., et al., *Dietary fiber and probiotics influence the gut microbiome and melanoma immunotherapy response*. Science, 2021. **374**(6575): p. 1632-1640.
268. Tanoue, T., et al., *A defined commensal consortium elicits CD8 T cells and anti-cancer immunity*. Nature, 2019. **565**(7741): p. 600-605.
269. Wolchok, J.D., et al., *Nivolumab plus ipilimumab in advanced melanoma*. N Engl J Med, 2013. **369**(2): p. 122-33.
270. Findley, K., et al., *Topographic diversity of fungal and bacterial communities in human skin*. Nature, 2013. **498**(7454): p. 367-70.
271. Hoarau, G., et al., *Bacteriome and Mycobiome Interactions Underscore Microbial Dysbiosis in Familial Crohn's Disease*. MBio, 2016. **7**(5): p. e01250-16.
272. Leonardi, I., et al., *Fungal Trans-kingdom Dynamics Linked to Responsiveness to Fecal Microbiota Transplantation (FMT) Therapy in Ulcerative Colitis*. Cell Host Microbe, 2020. **27**(5): p. 823-829 e3.
273. Lewis, J.D., et al., *Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease*. Cell Host Microbe, 2015. **18**(4): p. 489-500.
274. Liguori, G., et al., *Fungal Dysbiosis in Mucosa-associated Microbiota of Crohn's Disease Patients*. J Crohns Colitis, 2016. **10**(3): p. 296-305.

275. Sokol, H., et al., *Fungal microbiota dysbiosis in IBD*. Gut, 2017. **66**(6): p. 1039-1048.
276. Zhai, B., et al., *High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis*. Nat Med, 2020.
277. Zuo, T., et al., *Gut fungal dysbiosis correlates with reduced efficacy of fecal microbiota transplantation in Clostridium difficile infection*. Nat Commun, 2018. **9**(1): p. 3663.
278. Doron, I., et al., *Mycobiota-induced IgA antibodies regulate fungal commensalism in the gut and are dysregulated in Crohn's disease*. Nat Microbiol, 2021. **6**(12): p. 1493-1504.
279. Byrd, A.L., Y. Belkaid, and J.A. Segre, *The human skin microbiome*. Nat Rev Microbiol, 2018. **16**(3): p. 143-155.
280. Knutsen, A.P., et al., *Fungi and allergic lower respiratory tract diseases*. J Allergy Clin Immunol, 2012. **129**(2): p. 280-91; quiz 292-3.
281. Bradford, L.L. and J. Ravel, *The vaginal mycobiome: A contemporary perspective on fungi in women's health and diseases*. Virulence, 2017. **8**(3): p. 342-351.
282. Bongomin, F., et al., *Global and Multi-National Prevalence of Fungal Diseases-Estimate Precision*. J Fungi (Basel), 2017. **3**(4).
283. Brown, G.D., et al., *Hidden killers: human fungal infections*. Sci Transl Med, 2012. **4**(165): p. 165rv13.
284. Huffnagle, G.B. and M.C. Noverr, *The emerging world of the fungal microbiome*. Trends Microbiol, 2013. **21**(7): p. 334-41.
285. Vogtmann, E. and J.J. Goedert, *Epidemiologic studies of the human microbiome and cancer*. Br J Cancer, 2016. **114**(3): p. 237-42.
286. Helmink, B.A., et al., *The microbiome, cancer, and cancer therapy*. Nat Med, 2019. **25**(3): p. 377-388.
287. Sears, C.L., A.L. Geis, and F. Housseau, *Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis*. J Clin Invest, 2014. **124**(10): p. 4166-72.
288. Jain, T., et al., *New Insights Into the Cancer-Microbiome-Immune Axis: Decrypting a Decade of Discoveries*. Front Immunol, 2021. **12**: p. 622064.

289. Francescone, R., V. Hou, and S.I. Grivennikov, *Microbiome, inflammation, and cancer*. *Cancer J*, 2014. **20**(3): p. 181-9.
290. Leonardi, I., et al., *Mucosal fungi promote gut barrier function and social behavior via Type 17 immunity*. *Cell*, 2022. **185**(5): p. 831-846 e14.
291. Coker, O.O., et al., *Enteric fungal microbiota dysbiosis and ecological alterations in colorectal cancer*. *Gut*, 2018.
292. Liu, N.N., et al., *Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts*. *Nat Microbiol*, 2022. **7**(2): p. 238-250.
293. Yang, C.S., *Research on esophageal cancer in China: a review*. *Cancer Res*, 1980. **40**(8 Pt 1): p. 2633-44.
294. Chang, F., et al., *Infectious agents in the etiology of esophageal cancer*. *Gastroenterology*, 1992. **103**(4): p. 1336-48.
295. Malik, A., et al., *SYK-CARD9 Signaling Axis Promotes Gut Fungi-Mediated Inflammasome Activation to Restrict Colitis and Colon Cancer*. *Immunity*, 2018. **49**(3): p. 515-530 e5.
296. Wang, T., et al., *The Adaptor Protein CARD9 Protects against Colon Cancer by Restricting Mycobiota-Mediated Expansion of Myeloid-Derived Suppressor Cells*. *Immunity*, 2018. **49**(3): p. 504-514 e4.
297. Alam, A., et al., *Fungal mycobiome drives IL-33 secretion and type 2 immunity in pancreatic cancer*. *Cancer Cell*, 2022. **40**(2): p. 153-167 e11.
298. Dohlman, A.B., et al., *The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants*. *Cell Host Microbe*, 2020.
299. Walker, M.A., et al., *GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts*. *Bioinformatics*, 2018. **34**(24): p. 4287-4289.
300. Nash, A.K., et al., *The gut mycobiome of the Human Microbiome Project healthy cohort*. *Microbiome*, 2017. **5**(1): p. 153.

301. Proctor, D.M., et al., *Integrated genomic, epidemiologic investigation of Candida auris skin colonization in a skilled nursing facility*. Nat Med, 2021.
302. Zuo, T., et al., *Alterations in Fecal Fungal Microbiome of Patients With COVID-19 During Time of Hospitalization until Discharge*. Gastroenterology, 2020. **159**(4): p. 1302-1310 e5.
303. Ye, S.H., et al., *Benchmarking Metagenomics Tools for Taxonomic Classification*. Cell, 2019. **178**(4): p. 779-794.
304. Delsuc, F., H. Brinkmann, and H. Philippe, *Phylogenomics and the reconstruction of the tree of life*. Nat Rev Genet, 2005. **6**(5): p. 361-75.
305. Findley, K., et al., *Topographic diversity of fungal and bacterial communities in human skin*. Nature, 2013. **498**(7454): p. 367-70.
306. Saheb Kashaf, S., et al., *Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions*. Nat Microbiol, 2022. **7**(1): p. 169-179.
307. Brown, E.M., et al., *Phylogenetic analysis reveals a cryptic species Blastomyces gilchristii, sp. nov. within the human pathogenic fungus Blastomyces dermatitidis*. PLoS One, 2013. **8**(3): p. e59237.
308. Silverman, D., et al., *Skin involvement and breast cancer: are T4b lesions of all sizes created equal?* J Am Coll Surg, 2014. **219**(3): p. 534-44.
309. Commichaux, S., et al., *TaxaTarget: Fast, Sensitive, and Precise Classification of Microeukaryotes in Metagenomic Data*. 2021, Research Square.
310. Sifrim, D., et al., *Gastro-oesophageal reflux monitoring: review and consensus report on detection and definitions of acid, non-acid, and gas reflux*. Gut, 2004. **53**(7): p. 1024-31.
311. Li, X.V., et al., *Immune regulation by fungal strain diversity in inflammatory bowel disease*. Nature, 2022. **603**(7902): p. 672-678.
312. Fiers, W.D., I.H. Gao, and I.D. Iliev, *Gut mycobiota under scrutiny: fungal symbionts or environmental transients?* Curr Opin Microbiol, 2019. **50**: p. 79-86.
313. Limon, J.J., J.H. Skalski, and D.M. Underhill, *Commensal Fungi in Health and Disease*. Cell Host Microbe, 2017. **22**(2): p. 156-165.

314. Kumamoto, C.A., M.S. Gresnigt, and B. Hube, *The gut, the bad and the harmless: Candida albicans as a commensal and opportunistic pathogen in the intestine*. *Curr Opin Microbiol*, 2020. **56**: p. 7-15.
315. Aggor, F.E.Y., et al., *Oral epithelial IL-22/STAT3 signaling licenses IL-17-mediated immunity to oral mucosal candidiasis*. *Sci Immunol*, 2020. **5**(48).
316. Break, T.J., et al., *Aberrant type 1 immunity drives susceptibility to mucosal fungal infections*. *Science*, 2021. **371**(6526).
317. Fan, D., et al., *Activation of HIF-1alpha and LL-37 by commensal bacteria inhibits Candida albicans colonization*. *Nat Med*, 2015. **21**(7): p. 808-14.
318. David, L.A., et al., *Diet rapidly and reproducibly alters the human gut microbiome*. *Nature*, 2014. **505**(7484): p. 559-63.
319. Benedict, K., et al., *Epidemiologic and Ecologic Features of Blastomycosis: A Review*. *Current Fungal Infection Reports*, 2012. **6**(4): p. 327-335.
320. Dohlman, A.B. and X. Shen, *Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference*. *Exp Biol Med (Maywood)*, 2019. **244**(6): p. 445-458.
321. Ballou, E.R., et al., *Lactate signalling regulates fungal beta-glucan masking and immune evasion*. *Nat Microbiol*, 2016. **2**: p. 16238.
322. MacAlpine, J., et al., *A small molecule produced by Lactobacillus species blocks Candida albicans filamentation by inhibiting a DYRK1-family kinase*. *Nat Commun*, 2021. **12**(1): p. 6151.
323. Zeise, K.D., R.J. Woods, and G.B. Huffnagle, *Interplay between Candida albicans and Lactic Acid Bacteria in the Gastrointestinal Tract: Impact on Colonization Resistance, Microbial Carriage, Opportunistic Infection, and Host Immunity*. *Clin Microbiol Rev*, 2021. **34**(4): p. e0032320.
324. Ewaschuk, J.B., et al., *Secreted bioactive factors from Bifidobacterium infantis enhance epithelial cell barrier function*. *Am J Physiol Gastrointest Liver Physiol*, 2008. **295**(5): p. G1025-34.

325. Suerbaum, S. and P. Michetti, *Helicobacter pylori* infection. *N Engl J Med*, 2002. **347**(15): p. 1175-86.
326. Tang, Y.L., et al., *Detection and location of Helicobacter pylori in human gastric carcinomas*. *World J Gastroenterol*, 2005. **11**(9): p. 1387-91.
327. Nagao-Kitamoto, H. and N. Kamada, *Host-microbial Cross-talk in Inflammatory Bowel Disease*. *Immune Netw*, 2017. **17**(1): p. 1-12.
328. Ze, X., et al., *Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon*. *ISME J*, 2012. **6**(8): p. 1535-43.
329. Zhai, R., et al., *Strain-Specific Anti-inflammatory Properties of Two Akkermansia muciniphila Strains on Chronic Colitis in Mice*. *Front Cell Infect Microbiol*, 2019. **9**: p. 239.
330. Everard, A., et al., *Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity*. *Proc Natl Acad Sci U S A*, 2013. **110**(22): p. 9066-71.
331. Daillere, R., et al., *Enterococcus hirae and Barnesiella intestinihominis Facilitate Cyclophosphamide-Induced Therapeutic Immunomodulatory Effects*. *Immunity*, 2016. **45**(4): p. 931-943.
332. Aitchison, J., *The Statistical Analysis of Compositional Data*. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1982. **44**(2): p. 139-160.
333. Gloor, G.B., et al., *Microbiome Datasets Are Compositional: And This Is Not Optional*. *Front Microbiol*, 2017. **8**: p. 2224.
334. Basmaciyani, L., et al., *"Candida Albicans Interactions With The Host: Crossing The Intestinal Epithelial Barrier"*. *Tissue Barriers*, 2019. **7**(2): p. 1612661.
335. Chehoud, C., et al., *Fungal Signature in the Gut Microbiota of Pediatric Patients With Inflammatory Bowel Disease*. *Inflamm Bowel Dis*, 2015. **21**(8): p. 1948-56.
336. Banerjee, S., K. Schlaeppi, and M.G.A. van der Heijden, *Keystone taxa as drivers of microbiome structure and functioning*. *Nat Rev Microbiol*, 2018. **16**(9): p. 567-576.

337. Soler, A.P., et al., *Increased tight junctional permeability is associated with the development of colon cancer*. *Carcinogenesis*, 1999. **20**(8): p. 1425-31.
338. Ricciardi, M., et al., *Epithelial-to-mesenchymal transition (EMT) induced by inflammatory priming elicits mesenchymal stromal cell-like immune-modulatory properties in cancer cells*. *Br J Cancer*, 2015. **112**(6): p. 1067-75.
339. Ramirez-Garcia, A., et al., *Candida albicans and cancer: Can this yeast induce cancer development or progression?* *Crit Rev Microbiol*, 2016. **42**(2): p. 181-93.
340. Jawhara, S., et al., *Colonization of mice by Candida albicans is promoted by chemically induced colitis and augments inflammatory responses through galectin-3*. *J Infect Dis*, 2008. **197**(7): p. 972-80.
341. de Klerk, N., et al., *Lactobacilli Reduce Helicobacter pylori Attachment to Host Gastric Epithelial Cells by Inhibiting Adhesion Gene Expression*. *Infect Immun*, 2016. **84**(5): p. 1526-1535.
342. Zangl, I., et al., *The role of Lactobacillus species in the control of Candida via biotrophic interactions*. *Microb Cell*, 2019. **7**(1): p. 1-14.
343. Tjalsma, H., et al., *A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects*. *Nat Rev Microbiol*, 2012. **10**(8): p. 575-82.
344. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. *Bioinformatics*, 2009. **25**(14): p. 1754-60.
345. Ramirez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis*. *Nucleic Acids Res*, 2016. **44**(W1): p. W160-5.
346. Ramirez, F., et al., *High-resolution TADs reveal DNA sequences underlying genome organization in flies*. *Nat Commun*, 2018. **9**(1): p. 189.
347. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. *Nucleic Acids Res*, 2000. **28**(1): p. 27-30.
348. Breiman, L., *Random Forests*. *Machine Learning*, 2001. **45**(1): p. 5-32.
349. Abraham, A., et al., *Machine learning for neuroimaging with scikit-learn*. *Front Neuroinform*, 2014. **8**: p. 14.

350. Dhariwal, A., et al., *MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data*. *Nucleic Acids Res*, 2017. **45**(W1): p. W180-W188.
351. Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. *The Gene Ontology Consortium*. *Nat Genet*, 2000. **25**(1): p. 25-9.
352. Gene Ontology, C., *The Gene Ontology resource: enriching a GOld mine*. *Nucleic Acids Res*, 2021. **49**(D1): p. D325-D334.
353. Kanehisa, M., *Toward understanding the origin and evolution of cellular organisms*. *Protein Sci*, 2019. **28**(11): p. 1947-1951.
354. Martens, M., et al., *WikiPathways: connecting communities*. *Nucleic Acids Res*, 2021. **49**(D1): p. D613-D621.
355. Jassal, B., et al., *The reactome pathway knowledgebase*. *Nucleic Acids Res*, 2020. **48**(D1): p. D498-D503.
356. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0*. *Bioinformatics*, 2011. **27**(12): p. 1739-40.
357. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*. *Nucleic Acids Res*, 2009. **37**(1): p. 1-13.
358. Huang, D.W., et al., *DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists*. *Nucleic Acids Res*, 2007. **35**(Web Server issue): p. W169-75.
359. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool*. *BMC Bioinformatics*, 2013. **14**: p. 128.
360. Page, M.J., et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. *Rev Esp Cardiol (Engl Ed)*, 2021. **74**(9): p. 790-799.
361. Zhulin, I.B., *Databases for Microbiologists*. *J Bacteriol*, 2015. **197**(15): p. 2458-67.
362. Stulberg, E., et al., *An assessment of US microbiome research*. *Nat Microbiol*, 2016. **1**: p. 15015.

363. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. *Genome Biol*, 2014. **15**(12): p. 550.
364. Mukherjee, S., et al., *Genomes OnLine Database (GOLD) v.8: overview and updates*. *Nucleic Acids Res*, 2021. **49**(D1): p. D723-D733.
365. Kinane, D.F., P.G. Stathopoulou, and P.N. Papapanou, *Periodontal diseases*. *Nat Rev Dis Primers*, 2017. **3**: p. 17038.
366. Hajishengallis, G., *Periodontitis: from microbial immune subversion to systemic inflammation*. *Nat Rev Immunol*, 2015. **15**(1): p. 30-44.
367. Hu, J.M., et al., *Risk of colorectal cancer in patients with periodontal disease severity: a nationwide, population-based cohort study*. *Int J Colorectal Dis*, 2018. **33**(3): p. 349-352.
368. Michaud, D.S., et al., *Periodontal Disease Assessed Using Clinical Dental Measurements and Cancer Risk in the ARIC Study*. *J Natl Cancer Inst*, 2018. **110**(8): p. 843-854.
369. Visek, W.J., *Diet and cell growth modulation by ammonia*. *Am J Clin Nutr*, 1978. **31**(10 Suppl): p. S216-S220.
370. Lin, H.C. and W.J. Visek, *Colon mucosal cell damage by ammonia in rats*. *J Nutr*, 1991. **121**(6): p. 887-93.
371. Lin, H.C. and W.J. Visek, *Large intestinal pH and ammonia in rats: dietary fat and protein interactions*. *J Nutr*, 1991. **121**(6): p. 832-43.
372. Niederman, R., et al., *Ammonia as a potential mediator of adult human periodontal infection: inhibition of neutrophil function*. *Arch Oral Biol*, 1990. **35** Suppl: p. 205S-209S.
373. Lee, J.S., et al., *Urea Cycle Dysregulation Generates Clinically Relevant Genomic and Biochemical Signatures*. *Cell*, 2018. **174**(6): p. 1559-1570 e22.
374. Aristotle, *On the gait of animals*.
375. Burkholder, W.H., *Bacteria as plant pathogens*. *Annu Rev Microbiol*, 1948. **2** (1 vol.): p. 389-412.

376. McGovern, P.E., et al., *Fermented beverages of pre- and proto-historic China*. Proc Natl Acad Sci U S A, 2004. **101**(51): p. 17593-8.
377. Cao, Y., et al., *A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies*. Front Microbiol, 2017. **8**: p. 1829.
378. Lee, C.S., et al., *The Microbiota of Recreational Freshwaters and the Implications for Environmental and Public Health*. Front Microbiol, 2016. **7**: p. 1826.
379. Wagner, M. and A. Loy, *Bacterial community composition and function in sewage treatment systems*. Curr Opin Biotechnol, 2002. **13**(3): p. 218-27.
380. Bauer, M., et al., *The ALPK1/TIFA/NF-kappaB axis links a bacterial carcinogen to R-loop-induced replication stress*. Nat Commun, 2020. **11**(1): p. 5117.
381. Du, K., et al., *Construction of a gut microbiota-gene-pathway network to reveal the molecular mechanisms underlying right- and left-sided colorectal cancer*. FEMS Microbiol Lett, 2021. **368**(21-24).
382. Wang, J., et al., *Global Analysis of Microbiota Signatures in Four Major Types of Gastrointestinal Cancer*. Front Oncol, 2021. **11**: p. 685641.
383. Artola-Boran, M., et al., *Mycobacterial infection aggravates Helicobacter pylori-induced gastric preneoplastic pathology by redirection of de novo induced Treg cells*. Cell Rep, 2022. **38**(6): p. 110359.
384. Peuker, K., et al., *Microbiota-dependent activation of the myeloid calcineurin-NFAT pathway inhibits B7H3- and B7H4-dependent anti-tumor immunity in colorectal cancer*. Immunity, 2022. **55**(4): p. 701-717 e7.
385. Zhou, C.B., et al., *Fecal Signatures of Streptococcus anginosus and Streptococcus constellatus for Noninvasive Screening and Early Warning of Gastric Cancer*. Gastroenterology, 2022. **162**(7): p. 1933-1947 e18.

Biography

Anders B. Dohlman grew up in Chapel Hill, North Carolina with his parents Henrik Dohlman and Christianna Williams as well as his younger sister Camilla. He graduated from Carrboro High School in 2011 and Wesleyan University in 2015, where he received his B.A. in Mathematics and Biology. Prior to his doctoral research, Anders was mentored by Dr. Norman “Ned” Sharpless at the University of North Carolina, Drs. Peter Sorger and William Chen at Harvard Medical School, and Dr. Avi Ma’ayan at The Mount Sinai School of Medicine. For his post-doctoral work, Anders joined the lab of Dr. Matthew Meyerson at the Dana Farber Cancer Institute of Harvard Medical School.

His selected publications include:

Dohlman, A.B. and X. Shen, Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference. *Exp Biol Med* (Maywood), 2019. **244**(6): p. 445-458.

Dohlman, A.B., et al., The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe*, 2021. **29**(2): p. 281-298 e5.

Dohlman, A.B., et al., A pan-cancer mycobiome analysis reveals fungal involvement in gastrointestinal and lung tumors that is predictive of survival. *Under review*.

Dohlman, A.B., Zheng, J. et al., Taxonomic set enrichment analysis: a curated database and analytical toolkit for interpreting metagenomic data. *In preparation*.