Methodological Developments for the Analysis of Biological Samples in the Presence of Compositional Effects

By

Richard Meier

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Devin C Koestler, Committee Chair

Byron J Gajewski

Jeffrey Thompson

Prabhakar Chalise

Mary A Markiewicz

Date defended: February 18, 2020

The Dissertation Committee for Richard Meier certifies
that this is the approved version of the following dissertation:

Methodological Developments for the Analysis of Biological Samples in the Presence of Compositional Effects

Devin C Koestler, Committee Chair

Jo A Wick, Program Director

Date approved: February 28, 2020

Abstract

Compositional data, in which a vector of observed variables are constrained by a sum total, imposes a unique correlation structure among its components. Considering the abundance of components in a biological sample is inherently compositional, that is to say it is constrained by the amount of collected biomass, it is not surprising that compositional data frequently arises in the field of biomedical research. Failing to account for compositional effects compromises statistical inference, and may lead to spurious results that are not reproducible. Development and application of statistical techniques that honor compositionality in the context of biomedical research is therefore of great importance.

In this dissertation, we first investigate microbial composition in the pancreatic microbiome (not well characterized prior to this research) and surrounding tissue using a variety of different statistical methods. We identify similarities between tissue types and differences between tissues from subjects with different types of pancreatic cancer and tissues from non-cancer subjects. Identification of microbes commonly found in oral cavities then motivates the question whether consistent patterns of the microbial landscape with respect to disease can be found between the mouth and the gut.

Since there is no established method to test for these patterns in microbiome data, we continue by presenting a suitable Bayesian testing framework that is able to address the unique challenges posed by microbial abundance data. We elaborate how the method simultaneously applies to a variety of different data models and different types of estimates of microbial abundance, and demonstrate its ability to detect desired associations via simulation studies. Further, analysis of microbiome profiles derived from gut and oral cavity samples collected from pancreatic cancer cases are used to successfully identify microbes that exhibit consistent patterns of interest.

This dissertation closes with methodological developments for the analysis of DNA methylation levels of bulk samples with heterogeneous cell composition. Building on novel modelling approaches that are able to detect cell type specific methylation based on bulk samples, we introduce a Bayesian hierarchical modelling strategy that leverages spatial correlation of proximal CpG dinucleotides. We elaborate how our method was empirically motivated by whole blood methylation data of isolated cell types and demonstrate its performance improvement in terms of prediction accuracy and statistical power compared to non-spatial models.

Acknowledgements

First and foremost, I would like to thank my parents who never stopped loving and supporting me on my winding path through life; not even when I decided to move to the other side of the earth. I want to thank them for always putting their children first, for never failing to encourage us and for teaching us more than I could ever come close to writing down.

I want to thank my brothers for watching out for me, for teaching me kindness, for teaching me how to fight for myself, for challenging me to become a better person and for putting up with me for all these years.

I would like to thank my grandparents, who always worked hard through many adversities to raise and support their children and grandchildren. Your love will always be the cornerstone of our family and I miss you dearly.

I want to praise and thank my wife, Guangyi. Without her unending love and support this dissertation would not have been possible. She is my closest friend, my biggest cheerleader and my guiding compass. Thank you for being in my life and inspiring me to become the best version of myself every day.

Thank you to all my family and friends, who make my life worth living.

Special thanks goes to the Statistical Omics Working Group who have persistently provided me with helpful advise and feedback.

I also want to extend my sincere appreciation to my dissertation committee, Dr. Jeffrey Thompson, Dr. Byron Gajewski, Dr. Prabhakar Chalise, and Dr. Mary Markiewicz. I want to say "thank you" for the valuable time, the support and the revealing insights that they have provided me.

Last but by no means least, I have to express my most sincere gratitude to my advisor and dissertation committee chair, Dr. Devin Koestler. Throughout the entirety of my

journey at the University of Kansas Medical Center, he has been a relentless source of positivity, encouragement and guidance. His continuous support, his passion for my research and his deep insight were integral to the progress of my dissertation. It has been a true privilege to work with such an outstanding mentor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Compositional data traditionally refers to collections of non-negative random vectors in which each vector sums to a fixed (often arbitrary) constant and in which each component of a random vector contains relative information (Quinn et al., 2018), i.e. information that is carried by ratios of vector components and not by the values of components themselves (Pawlowsky-Glahn & Egozcue, 2016). Compositional data is encountered in many research settings across a wide variety of fields, and most commonly arises when the data of interest can be expressed as proportions of a total. Failing to acknowledge the constant-sum constraint that underlies compositional data can have serious implications for experiments that aim to assess how individual components of the random vectors associate. Negative correlations among components of compositional random vectors are favored and will necessarily arise (Chayes, 1960), since an increase in one component necessarily causes the decrease of other components. Spurious associations frequently arise, especially when trying to subset or aggregate the data, and classical methods that assume independence run the risk of assigning significance to false-positive, non-reproducible associations that are solely a product of the constraints and of which components are included into the analysis (Gloor et al., 2017; Quinn et al., 2019).

In biomedical analyses, compositional data often emerge because biological landscapes are inherently heterogeneous, comprising complex ensembles of interacting molecules, cells and functional structures, which all sum to the total biomass collected in a given sample. This issue is compounded by the fact that in the research of severe diseases that are challenging to effectively cure, such as many forms of cancer, mechanisms on the cellular level and their connection to disease are evaluated especially frequently. An intuitive example that has received a lot of attention in recent years due to its association with many diseases, is the analysis of microbial abundance (Tsilimigras & Fodor, 2016; Gloor et al., 2017). In microbial abundance data, the number of times that a microbial species is observed in a given biological sample is limited by the total number of

microbes either observed or contained in that sample. On the other hand, one of the most influential and powerful set of tools to study cellular processes, next generation sequencing techniques, are inherently based on an array of counts whose magnitude is dictated by the sequencing depth of samples (Quinn et al., 2018, 2019)). Even if the primary outcome of a target study is not compositional in nature, outcomes affecting covariates that are themselves compositional can complicate the analysis. Ongoing research in the field of compositional data analysis (Martino et al., 2019; Quinn et al., 2019; Hawinkel et al., 2019; Ishiya & Aburatani, 2019)) shows that it is crucial to consider development and application of statistical methods that acknowledge compositionality.

In Chapter 2, the overall microbiome in pancreatic tissue, which was not well known prior to this research, is characterized. Differences and similarities in microbial composition between gut tissues of pancreatic cancer and non-cancer subjects are analyzed by employing multivariate, distance based approaches, as well as via univariate Beta regression that explores groups of microbial species individually. Microbes for which abundance shows a significant difference between cancer and non-cancer subjects after adjusting for potential confounders are successfully identified. Regardless of disease status, the microbial landscape of biological samples was found to be highly subject specific while also exhibiting clear similarities between pancreatic and duodenum tissue. Both pancreatic and duodenum tissues contained, but were not limited to, microbes commonly identified in the oral cavity.

These results in combination with other studies that found associations between diseases of the gut and microbiomes in both gut and mouth motivate the question of whether there are microbes that show consistent patterns of association with respect to some phenotype of interest, between mouth and gut. Identification of such patterns could potentially provide information about the microbial composition in the gut based on the microbial composition of oral samples, which can be collected using minimally invasive techniques. Further, the identification of consistent patterns of microbial composition between the gut and oral cavity, two anatomically distinct sites, may offer new insights into the interrelatedness of the microbiome throughout the human body. While oral microbiome samples were collected on pancreatic cancer cases used in the first research project,

there remains a gap in the literature in terms of statistical methods designed to identify these types of patterns given the unique characteristics of microbiome data, e.g., compositional, high dimensional, sparse (zero-inflated).

In Chapter 3 we address this unmet need using a novel Bayesian framework that allows researchers to formally test whether the abundance of a group of microbes is associated between body sites with respect to some phenotypic variable. This method is built to be applicable to a variety of different statistical models, making it suitable to address the unique challenges posed by sparse compositional data. Viability of the approach is first demonstrated using a variety of simulation studies that consider different degrees of association and different data models. The chapter concludes by applying the approach to data derived from pancreatic cancer cases and successfully identifying microbes that exhibit consistent patterns between gut and mouth with respect to disease subtype.

Chapter 4 shifts the focus towards recently developed statistical methods for DNA methylation analyses that estimate cell-type specific methylation levels from bulk samples. While existing statistical methods adequately account for, and even take advantage of the underlying cell-composition of a given sample, they do not consider the well-recognized spatial correlation between genomically proximal CpG loci. In this chapter, a Bayesian hierarchical modelling strategy is presented that builds on existing methods and leverages spatial correlation in the methylation levels of nearby CpGs to improve the operating characteristics of statistical tests for differential methylation with respect to some condition or exposure. First, tuning parameters of the method are informed based on empirical evaluation of whole blood methylation data of isolated cell types. Next, extensive simulation studies are employed to both identify the most suitable hierarchical model among a set of candidate models, and to compare benefits of spatial models compared to non-spatial models. The results of our simulation studies demonstrated improved prediction accuracy and statistical power of the proposed spatial models as compared to models that ignore the spatial correlation in the methylation levels of nearby CpGs.

Chapter 2

The Microbiomes of Pancreatic and Duodenum Tissue Overlap and are Highly Subject Specific but Differ between Pancreatic Cancer and Non-Cancer Subjects

This chapter has previously been published and is reprinted here with permission with minor adaptations. del Castillo, E.*, Meier, R.*, Chung, M., Koestler, D. C., Chen, T., Paster, B. J., Charpentier, K. P., Kelsey, K.T., Izard, J., Michaud, D. S. (2019). The Microbiomes of Pancreatic and Duodenum Tissue Overlap and are Highly Subject Specific but Differ between Pancreatic Cancer and Non-Cancer Subjects. *Cancer Epidemiol. Biomark. Prev.* 28(2):370-383.

* *del Castillo, E. and Meier, R. both share first authorship*

2.1 Statement of Contributions

In this project, I, Richard Meier, performed data cleaning and all statistical analyses that utilize zero-inflated Beta regression models. I also helped to interpret results and to write the original manuscript. Contributions of other authors are listed below.

KPC contributed to the identification of patients and collection of specimens. ED carried out the DNA extractions and preparation of sample for sequencing. TC conducted the taxonomic assignment of the 16S rRNA amplicon sequencing data. RM, MC, DCK performed the statistical analysis. TC, ED contributed to the QIIME analysis. DSM, JI, KPC contributed to the design and funding of the study. DSM, BJP, KTK participated in the coordination of the study. DSM, ED, RM helped write the manuscript. All authors were involved in the interpretation of results and editing of the final manuscript.

2.2 Abstract

Background: In mice, bacteria from the mouth can translocate to the pancreas and impact pancreatic cancer progression. In humans, oral bacteria associated with periodontal disease have been

linked to pancreatic cancer risk. It is not known if DNA bacterial profiles in the pancreas and duodenum are similar within individuals.

Methods: Tissue samples were obtained from 50 subjects with pancreatic cancer or other conditions requiring foregut surgery at the Rhode Island Hospital (RIH), and from 34 organs obtained from the National Disease Research Interchange. 16S rRNA gene sequencing was performed on 189 tissue samples (pancreatic duct, duodenum, pancreas), 57 swabs (bile duct, jejunum, stomach), and 12 stool samples.

Results: Pancreatic tissue samples from both sources (RIH and National Disease Research Interchange) had diverse bacterial DNA, including taxa typically identified in the oral cavity. Bacterial DNA across different sites in the pancreas and duodenum were highly subject specific in both cancer and noncancer subjects. Presence of genus Lactobacillus was significantly higher in noncancer subjects compared with cancer subjects and the relative abundance of Fusobacterium spp., previously associated with colorectal cancer, was higher in cancer subjects compared with noncancer subjects.

Conclusions: Bacterial DNA profiles in the pancreas were similar to those in the duodenum tissue of the same subjects, regardless of disease state, suggesting that bacteria may be migrating from the gut into the pancreas. Whether bacteria play a causal role in human pancreatic cancer needs to be further examined.

2.3 Introduction

In 2018, an estimated 55,440 individuals will be diagnosed with pancreatic cancer in the US, and only 8% of these individuals are expected to survive the next five years (ACS, 2018). Given this high fatality rate, and the silent progression of early disease, identifying risk factors for the prevention and early detection of pancreatic cancer is critical to reducing its mortality. To date, known risk factors for pancreatic cancer, including smoking, obesity, diabetes, heavy alcohol consumption, family history and markers of genetic susceptibility, cannot, even collectively, be used for early detection and risk stratification of pancreatic cancer in the general population (Klein et al.,

2013).

Studies have suggested a link between bacteria and pancreatic cancer risk (Michaud, 2013), highlighting the need to more critically explore the underlying factors that affect the microbiome of the oral cavity and upper digestive tract in both cancer patients and cancer-free individuals. The current research on oral bacteria and pancreatic cancer risk stems from a number of observational studies that reported a higher risk of pancreatic cancer among individuals with periodontitis, when compared to those without periodontitis (Michaud, 2013; Michaud et al., 2017). Periodontitis, an inflammatory disease of the gums, is largely driven by keystone pathogens and pathobionts (Hajishengallis, 2014). Two large prospective cohort studies have reported positive associations between periodontal disease pathogens and subsequent pancreatic cancer risk (Michaud et al., 2012a; Fan et al., 2016); in these two studies, detection of elevated antibodies to *Porphyromonas gingivalis*, measured in blood collected prior to cancer diagnosis, was associated with a two-fold higher risk of pancreatic cancer (Michaud et al., 2012a), and presence (vs absence) of *P. gingivalis* in saliva collected prior to cancer diagnosis was associated with a 60% increase in risk of pancreatic cancer (Fan et al., 2016). *Aggregatibacter actinomycetemcomitans*, another periodontal pathogen, was also associated with pancreatic cancer risk in the prospective study using saliva (Fan et al., 2016).

Few investigations to date have attempted to detect bacteria in pancreatic tissue. Earlier studies reported the presence of bacteria in pancreatic ducts of subjects with chronic pancreatitis or bile duct obstruction (Swidsinski, 2005; Schneider et al., 2015; Scheithauer et al., 2009). Other studies have investigated the presence of specific bacterial DNA in the pancreatic tissue of pancreatic cancer subjects, namely species of *Helicobacter* (Nilsson, 2006) and *Fusobacterium* (Mitsuhashi et al., 2015). The most comprehensive molecular microbiome studies to date reported the presence of a diverse bacterial populations in fluids collected from the bile duct, pancreas and jejunum of subjects undergoing pancreaticoduodenectomy (Rogers et al., 2017), and in pancreatic cyst fluid removed endoscopically from pancreatic cysts (Li et al., 2017). In mice, bacteria have been shown to translocate from the mouth to the pancreas, and germ-free mice have reduced progression of

pancreatic ductal adenocarcinoma (Pushalkar et al., 2018).

Metagenomics studies on DNA isolated from tissue samples from cancer subjects have been conducted for lung (Yu et al., 2016), colorectal (Bullman et al., 2017), esophageal (Elliott et al., 2017), stomach (Wang et al., 2016), and breast cancer (Hieken et al., 2016) These studies demonstrate that 16S rRNA gene sequencing can be effectively conducted on fresh tissue samples where the ratio of bacterial to human DNA is much lower than at other human sites (e.g., stool or oral cavity)(Segata et al., 2012). Moreover, these studies have shown that bacterial profiles at different organ sites are often unique (Yu et al., 2016) and that changes may be associated with cancer (Elliott et al., 2017; Wang et al., 2016). In two recent studies, bacterial DNA was measured in tumor tissue samples obtained from patients with pancreatic ductal adenocarcinoma (PDAC) using 16S rRNA gene sequencing (Pushalkar et al., 2018; Geller et al., 2017); however, comparison of microbiota in pancreas and different gastrointestinal tissue was not conducted in these patients.

To date, no study to our knowledge has characterized the overall microbiome in pancreatic and normal surrounding tissue samples, a critical step to understand whether and how bacteria may play a role in carcinogenesis. In an effort to address the specific question of whether the pancreas has its own microbiome, we recruited subjects from the Rhode Island Hospital (Providence, RI) with planned foregut surgery to obtain tissue samples for 16S rRNA gene microbiome analysis. In addition, for comparison to controls, we obtained pancreatic and duodenum tissue from National Disease Research Interchange (NDRI) from non-cancer subjects.

2.4 Materials and Methods

2.4.1 Study population and sample collection

Seventy-seven subjects, enrolled between January 2014 and March 2016, were included in this study. Subjects were eligible if identified as candidates for surgery of the foregut by Dr. Charpentier (the lead surgeon at the RIH) and included those with pancreatic cancer, pancreatic cystic neoplasms, pancreatitis, bile duct or small bowel diseases. All recruited subjects were between 31-86 years old (Table 2.1). Participants were asked to complete a self-administered questionnaire

to provide data on demographic and behavioral factors, and included a question on past use of antibiotics; this variable was included in the statistical analysis to control for changes that may have occurred due to antibiotic use in recent past. Questions on family history of cancer, use of other over-the-counter medications were also included on the questionnaire. Stool collection kits with ethanol as a fixative (95% (wt/wt) ethanol) were provided prior to surgery (Franzosa et al., 2014); participants were asked to return the samples using a pre-paid box.

A protocol was established for processing tissue samples collected during surgery to reduce contamination. A technician from the Pathology Department was informed in advance of the surgery date and time, and was paged as soon as the specimens had been obtained. Surgical tissue samples were frozen within one hour of the surgery time, as well as tissues swab samples from the stomach, jejunum, and bile duct that were collected using DNA-free forensic sterile swabs whenever possible. During surgery, the surgeon also recorded (on a surgery form for the study) if the patient had received prior pre-OP endoscopic ultrasound (EUS), had previously had their gallbladder removed, or had received prior placement of a stent (for treatment of symptoms); all subjects received a single dose of perioperative antibiotics immediately prior skin incision at the time of the operation. Tissue samples (pancreatic tumor tissue, pancreatic cysts, normal pancreatic tissue, pancreatic ducts and duodenums) were prepared by a Rhode Island Hospital pathologist; cancerous and non-cancerous tissues were identified, separated and labeled. All samples were de-identified and stored at -80°C until processing.

Upon review of pathology records, ICD10 codes were assigned to each subject; 39 subjects had pancreatic cancer (ICD10 codes C25.0-C25.9; the majority of cases were adenocarcinomas, only 2 subjects had neuroendocrine tumors of the pancreas), 12 subjects had periampullary cancer (ICD10 codes C24.0-C24.1), 18 subjects had other pancreatic conditions (ICD10 codes K86.0-K86.3), and the remaining 8 had other gastrointestinal conditions. The study was approved by Lifespan's Research Protection Office for recruitment at RIH, as well as the Institutional Review Boards for Human Subjects Research at Brown University, Tufts University and the Forsyth Institute.

In addition, we obtained pancreatic specimens without known conditions of pancreatic dis-

Table 2.1: Distribution of demographic, lifestyle, and health conditions variables among patients with diseases of the foregut, primarily pancreatic diseases, and deceased controls

RIH subjects (n = 77)		NDRI subjects (n = 34)	
Characteristic	Mean (SD)	Characteristic	Mean (SD)
Age	63 ± 13	Age	68 ± 15
Body mass index	27 ± 6	Body mass index	29 ± 6.5
	N (%)		N (%)
Sex		Sex	
Male	38 (49)	Male	21 (62)
Female	39 (51)	Female	13 (38)
Race		Race	
Caucasian	72 (93.5)	Caucasian	30 (88)
Black	2 (2.6)	Black	2 (6)
Other	2 (2.6)	Other	2 (6)
Smoking status		Smoking status	
Ever smoker	44 (58)	Ever smoker	23 (68)
Chemotherapy		Cause of death	
Never	52 (76.5)	Heart failure	17 (50)
Prior to past 6 months	7 (10.3)	Cardiopulmonary arrest	5 (15)
In past 6 months	9 (13.2)	Cerebrovascular accident	1 (3)
		Respiratory arrest	2 (6)
Antibiotic use		Abdominal aortic aneurysm	1 (3)
Never	13 (18.1)	Intracerebral hemorrhage	1 (3)
Prior to past 6 months	32 (44.2)	Liver cirrhosis	1 (3)
In past 6 months	21 (29.2)	Overdose	1 (3)
Missing	6 (8.3)	Parkinson's disease	1 (3)
		Pneumonia	1 (3)
Stent prior to surgery (yes)	19	Pulmonary embolism	1 (3)
Pre-OP EUS	20	Pulmonary fibrosis	1 (3)
Surgery for:			
Pancreatic cancer	51 (66.2)		
Chronic pancreatitis or pancreatic cysts	18 (23.4)		
Other	8 (10.4)		

eases from the National Disease Research Interchange (NDRI) to serve as control samples in the absence of available healthy pancreatic tissue in non-cancer subjects. Snap-frozen 'control' whole-pancreas and duodenum ($\sim$ 5cm) human specimens from 34 deceased donors were obtained from NDRI with an average post-mortem recovery time of 13 hours. Control pancreas (head and tail), pancreatic ducts and duodenums were dissected under sterile conditions, and stored at -80°C until processing. To remove additional contamination, we removed a thin tissue layer around each sample prior to extracting DNA. Details for DNA extraction and sequencing procedures are provided in the Supplementary Methods.

2.4.2 16S rRNA amplicon Illumina sequencing

The 16S rRNA gene dataset consists of Illumina MiSeq sequences targeting the V3-V4 hypervariable regions. The DNA target sequencing was performed by the Forsyth Institute Sequencing Core. To evaluate effect of running samples on MiSeq runs at different times, we included bacterial mock community samples on each run and then compared their relative abundances across the MiSeq runs; the results for the mock communities were consistent across run, demonstrating minor fluctuations (Supplementary Figure A.1).

The MiSeq reporter analysis was used to discard low quality sequences and to generate FASTQ files containing only filtered quality sequences, subsequently the overlapping paired-end reads were stitched together and further processed using used a multi-stage BLASTN-base search taxonomy read assignment pipeline that maximizes species level classification (Al-Hebshi et al., 2015).

2.4.3 Taxonomic assignment pipeline of 16S rRNA amplicon sequencing data

Sequences were BLASTN-searched against a combined set of 16S rRNA reference sequences that consist of the HOMD (version 14.5)(Dewhirst et al., 2010), Greengenes Gold (McDonald et al., 2011), and the NCBI 16S rRNA reference sequence set. All assigned reads were subject to several down-stream bioinformatics analyses, including alpha and beta diversity assessments, provided in the QIIME (Quantitative Insights Into Microbial Ecology (Caporaso et al., 2010)) software package

version 1.9.1.

2.4.4 Statistical analysis

Samples with < 500 total read counts were excluded from all analysis. In addition, only OTUs with a minimal read count of 100 sequences (across all samples) were included in the analyses. For QIIME analyses, we normalized the number of sequences in the different MiSeq runs by rarefying each library to 500 reads to account for differences in sequencing depth across runs (increasing rarefaction cutpoint to higher read number did not result in changes in alpha-diversity results or OTU numbers in samples; 500 reads was used as the cutpoint to reduce number of samples lost from the analysis). Range of sequencing counts for the different sample types are provided in Supplementary Figure A.2.Across samples, OTU relative abundance was computed as the ratio of an OTU's absolute abundance to the total number of reads for that sample.

To create relative abundance plots, we restricted bacterial taxa (at genus-level) present at >2% relative abundance and with >35% prevalence in both NDRI and RIH samples (this was done to simplify comparison between the RIH and NDRI samples). Jaccard Index was used for paired comparison of proportion of shared microbiota taxa present at >2% relative abundance in tissue/swab samples within subjects.

To examine the variation in the microbial profile across the different habitats/sites (Supplemental Table A.1) among the NDRI and RIH subjects, we calculated the distance/dissimilarity between samples using the Bray-Curtis and Sorensen indices (Bray & Curtis, 1957). Computed distances were subsequently used to generate principal coordinate analysis (PCoA) plots to visualize the arrangement of the samples in the ordination space. PERMANOVA (available in QIIME) was used to test whether the distances are more similar within a group of samples than that from other groups of samples.

To identify demographic and clinical correlates of pancreatic microbial composition, we fit a series of zero-inflated beta regression models to examine associations between genus-level relative abundances and demographic (i.e., age, gender, race, BMI) and clinical (i.e., health sta-

tus, chemotherapy, antibiotics use prior to surgery, anxiety medications, presence of stent prior to surgery, whether pre-operative endoscopic ultrasound [pre-OP EUS] was conducted prior to surgery, tumor surgery classification by International Code of Disease [ICD10 code]). In the results, we refer to relative mean abundance among non-zero observations (μ) merely as relative mean abundance. More details are provided in the Supplemental Methods.

We explored which factors obtained in the questionnaires and medical files in RIH subjects were associated with bacterial communities in the pancreatic tissue samples. The most influential factors were sequencing run, presence of stent, and chemotherapy prior to surgery (only 5 patients with available tissue/swab samples had chemotherapy in the past 6 months); each of these factors was significantly associated with a large number of genera tested in marginal models. Given that the mock bacterial communities were similar across runs (see Supplemental Materials), it is possible that "run" was associated with certain genera due to differences in number of samples per sequencing run. To adjust for potential confounding, we considered this covariate in the final models comparing cancer to non-cancer subjects and the different ICD-codes among the RIH subjects. Similarly, age, BMI and sex were adjusted for as these features were shared between the studies and were found to explain variation in the relative abundance of some of the genera. Smoking was not found to explain variation in relative abundance in our data.

2.5 Results

The present analysis included a total of 246 pancreatic tissue and swab samples collected from 82 subjects (50 subjects from RIH providing 133 samples [57 swabs, 76 tissue] and 34 subjects from NDRI providing 113 tissue samples; Supplemental Table A.1). In addition, 12 RIH subjects provided stool samples. There were no significant differences in the distribution of age, gender, BMI, race, and smoking status between RIH and NDRI subjects (Table 2.1). The Illumina-based sequencing of V3-V4 hypervariable regions of the bacterial 16S rRNA gene resulted in a total of 19,498,743 high quality sequences (with a median sequence length of 427 nucleotides).

2.5.1 Taxonomy

Over 99% of the reads from RIH pancreatic samples were attributed to 5 bacterial phyla (45.9% Proteobacteria, 35.6% Firmicutes, 9.5% Bacteroidetes, 4.3% Fusobacteria, and 3.9% Actinobacteria). The remaining low abundance phylotypes (0.6% of the total) belonged to six bacterial phyla (Synergistetes, TM7, Deinococcus-Thermus, Verrucomicrobia, Spirochaetes, and Tenericutes). 99.6% of the reads observed among the NDRI pancreatic samples belonged to the same five bacterial phyla as observed in RIH subjects. The phylum Tenericutes (Bacteria) was present only in RIH samples, and the phylum Euryarchaeota (Archaea) was present only in NDRI samples, but both of these phyla were uncommon.

While the microbial communities in the pancreatic tissues were dominated by the phyla Firmicutes and Proteobacteria, substantial inter-individual variability was observed. In RIH samples, Proteobacteria relative abundance ranged from 2 to 99%, and similarly, Firmicutes relative abundance ranged from 0.6 to 84%. Large inter-individual variability was also observed in the NDRI samples.

2.5.2 Within and between sample diversity analysis

Mean relative abundance for bacterial taxa (mostly at the genus-level) in the pancreatic tissue samples (duct, head, tail, normal and tumor), duodenum tissue samples, and jejunum, bile duct and stomach swabs are presented for each subject with more than one available sample in the RIH in Figure 2.1 and NDRI in Figure 2.2 (males and females are presented separately for ease of comparison - no major differences were observed by sex). Three striking patterns emerge: 1) bacterial profiles in the pancreas are subject-specific rather than site-specific, 2) bacterial profiles in duodenum tissue are remarkably similar to those in pancreatic tissue in the same subjects, 3) concordance of paired comparisons of bacterial profiles in cancer subjects (RIH) are slightly lower across tissue type or site than those for non-cancer subjects (NDRI) (Figure 2.3). Subjects from RIH with only one sample available (n=5) demonstrate similar bacterial profiles as those with multiple samples. Bacterial taxa commonly recognized as oral bacteria, including *Fusobacterium spp.*,

Prevotella spp., *Dialister spp.*, *Veillonella spp.*, and *Haemophilus spp.* were identified in many of the tissue samples, both cancer and non-cancer subjects (Figure 2.1). Other oral bacterial taxa, including *Parvimonas micra*, *Selenomonas noxia*, *Capnocytophaga spp.*, *Peptostreptococcus spp.* and *Solobacterium moorei* were also identified in tissue samples but were less common (present in 20%-35% of all samples).

With the exception of the stool and the jejunum, all the bacterial communities were characterized as habitats with low bacterial richness including the pancreatic sites, duodenums and the bile ducts (Figure 2.4 A). Among RIH subjects, the microbial communities of the stool samples were represented by higher richness than the microbial communities in the tumors of the pancreas ($p = 0.007$), duodenums ($p = 0.013$) and bile duct swabs ($p = 0.017$). Likewise, the stool bacterial communities had higher richness than the NDRI pancreatic heads ($p = 0.012$), pancreatic ducts ($p = 0.020$) and duodenums ($p = 0.005$). The microbial communities in the jejunum swabs showed more richness than the communities in the RIH pancreatic head ($p = 0.014$) and duodenums ($p = 0.028$). In general, the bacterial communities in the pancreas of RIH subjects had slightly higher richness when compared to those from the pancreas of the NDRI matching sample types. Similar results were observed using additional alpha diversity measures of the bacterial communities (Figure 2.4 B-D). As expected, the stool samples were the most diverse with a Shannon index ≥ 4 (Figure 2.4 B).As the number of phyla represented in high abundance in these samples was relatively low (~ 5), we observed relatively low levels of phylogenetic distances across all samples (Figure 2.4 D).

The ordination beta-diversity analysis revealed that the majority of samples belonged to a single cluster, without any visually apparent groupings by the nature of the sample, health status or anatomical site (Figure 2.5 A-C). However, the PERMANOVA tests revealed statistically significant differences between NDRI and RIH samples ($p < 0.001$), and for the swab samples obtained from the bile duct, jejunum and stomach (compared to pancreas tissue samples). Differences between sites within the pancreas (i.e., head, tail, duct), and compared to the duodenum (for NDRI and RIH, separately), were not statistically significant (after accounting for multiple comparisons).

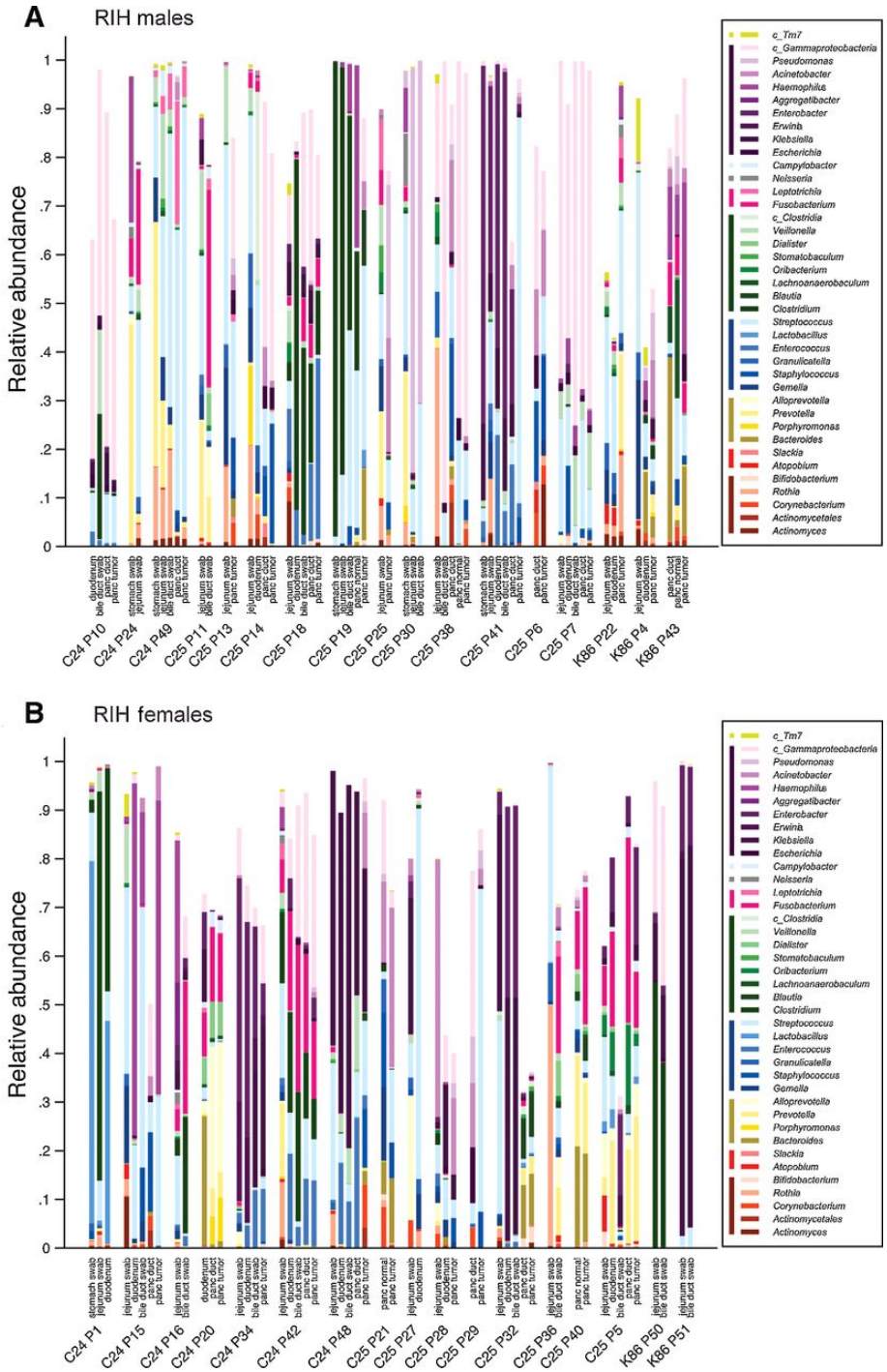


Figure 2.1: (A) RIH Males with ICD code (C25, C24 or K86); (B) RIH Females with ICD code (C25, C24, or K86). Distribution of bacteria relative abundance by genus level in all the studied body habitats based on read taxa attribution using V3-V4 hypervariable region of 16S rRNA genes. All names are at genera level except for those with c_ which denotes class for multi-genera taxa (within that class). Colored bars next to legend reflect taxa at class level: TM7 (lime); Gammaproteobacteria (purple); Epsilonproteobacteria (light grey); Betaproteobacteria (dark grey); Fusobacteriia (pink); Clostridia (green); Bacilli (blue); Bacteroides (Gold); Coriobacteriia (red); Actinobacteria (marron).

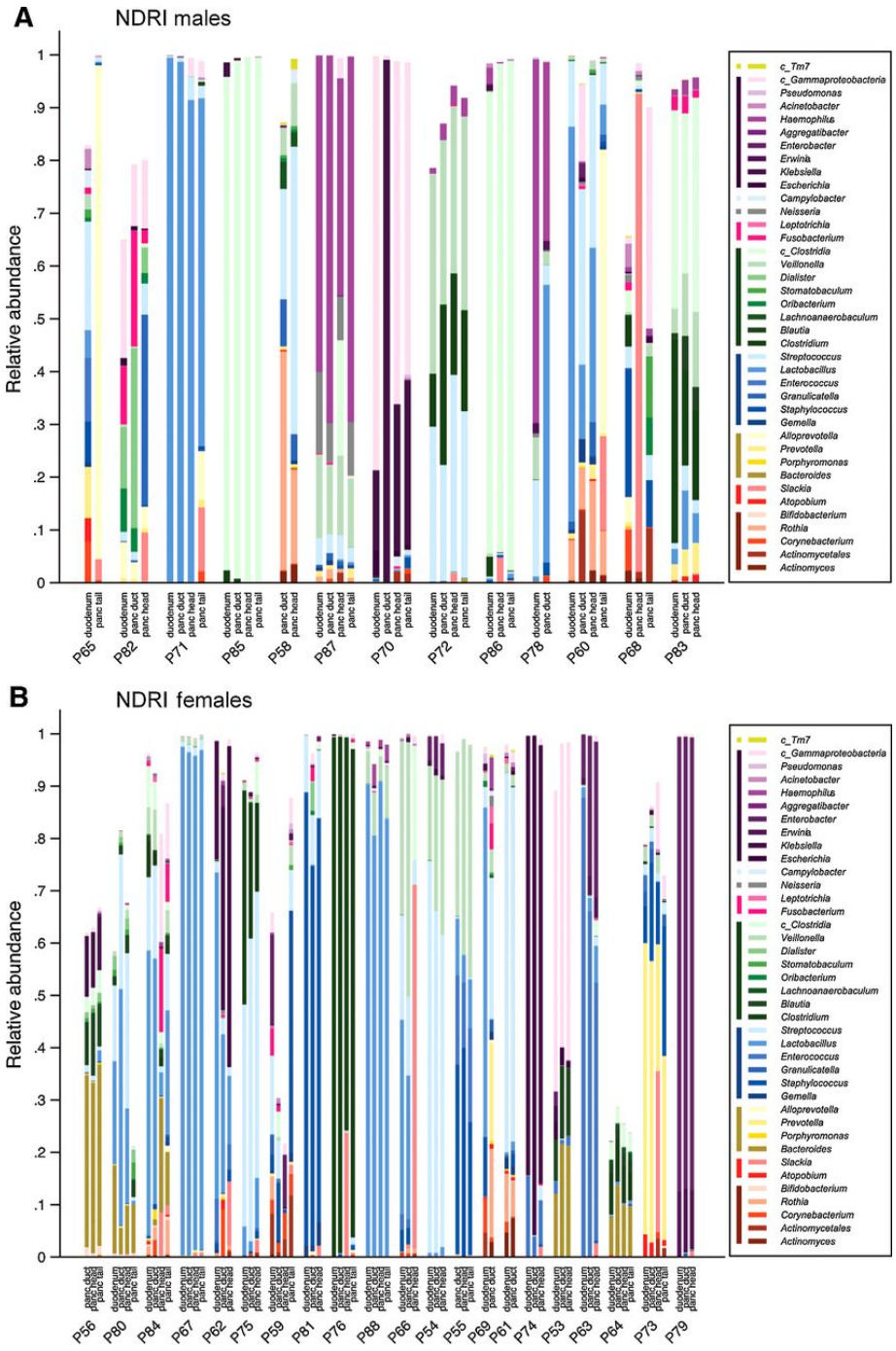


Figure 2.2: (A) NDRI Males; (B) NDRI Females. Distribution of bacteria relative abundance by genus level in all the studied body habitats based on read taxa attribution using V3-V4 hypervariable region of 16S rRNA genes. All names are at genera level except for those with c_ which denotes class for multigenera taxa (within that class). Colored bars next to legend reflect taxa at class level: TM7 (lime); Gammaproteobacteria (purple); Epsilonproteobacteria (light grey); Betaproteobacteria (dark grey); Fusobacteriia (pink); Clostridia (green); Bacilli (blue); Bacteroides (Gold); Coriobacteriia (red); Actinobacteria (marron).

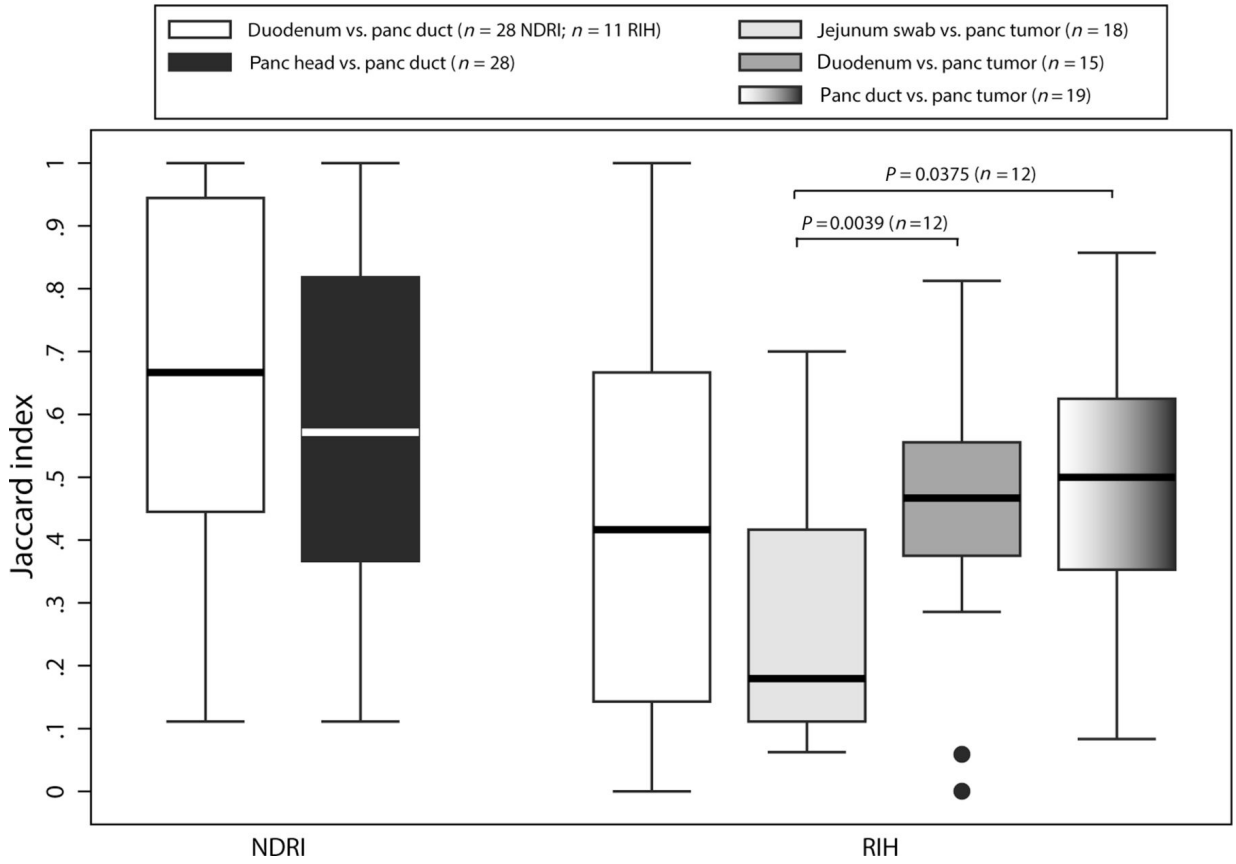


Figure 2.3: Jaccard Index (proportion of shared genera) for paired comparison of tissue samples in NDRI and RIH subjects.

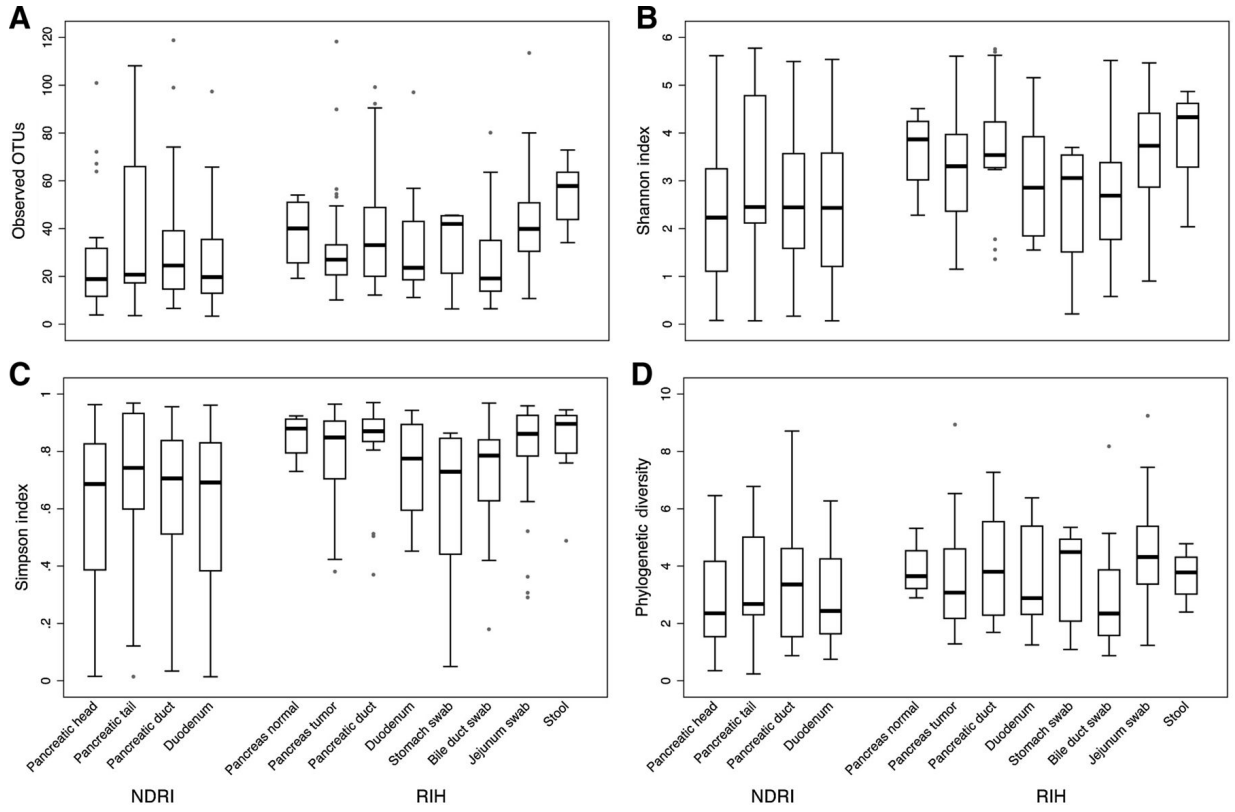


Figure 2.4: Comparative alpha diversity analyses of bacterial communities in anatomical sites (based on a simulated data set subsampled from the input OTU table). Alpha diversity metrics: (A) Richness, (B) Shannon diversity index, (C) Simpson index, and (D) Phylogenetic diversity.

The principal component analyses of both Bray-Curtis and Sorensen distances between all samples (tissues and swabs)showed that both RIH and NDRI samples clustered mostly by subject.

2.5.3 Associations of host factors with microbial communities

Using multiple regression analyses, we examined presence or absence, and relative mean abundance of bacterial taxa (at the genus and species level) among present (non-zero) observations using all tissue and swab samples comparing RIH subjects to NDRI subjects. Table 2.2 presents the bacterial taxa (at the genus level) that were present in at least 20% of all tissue and swab samples; each model includes both a zero-inflated component (testing for differences in presence/absence of bacterial taxa) and a relative mean abundance comparison. *Lactobacillus* taxa were present in almost all non-cancer tissue samples (estimated proportion of presence $[P1] = 0.98$), but were much less

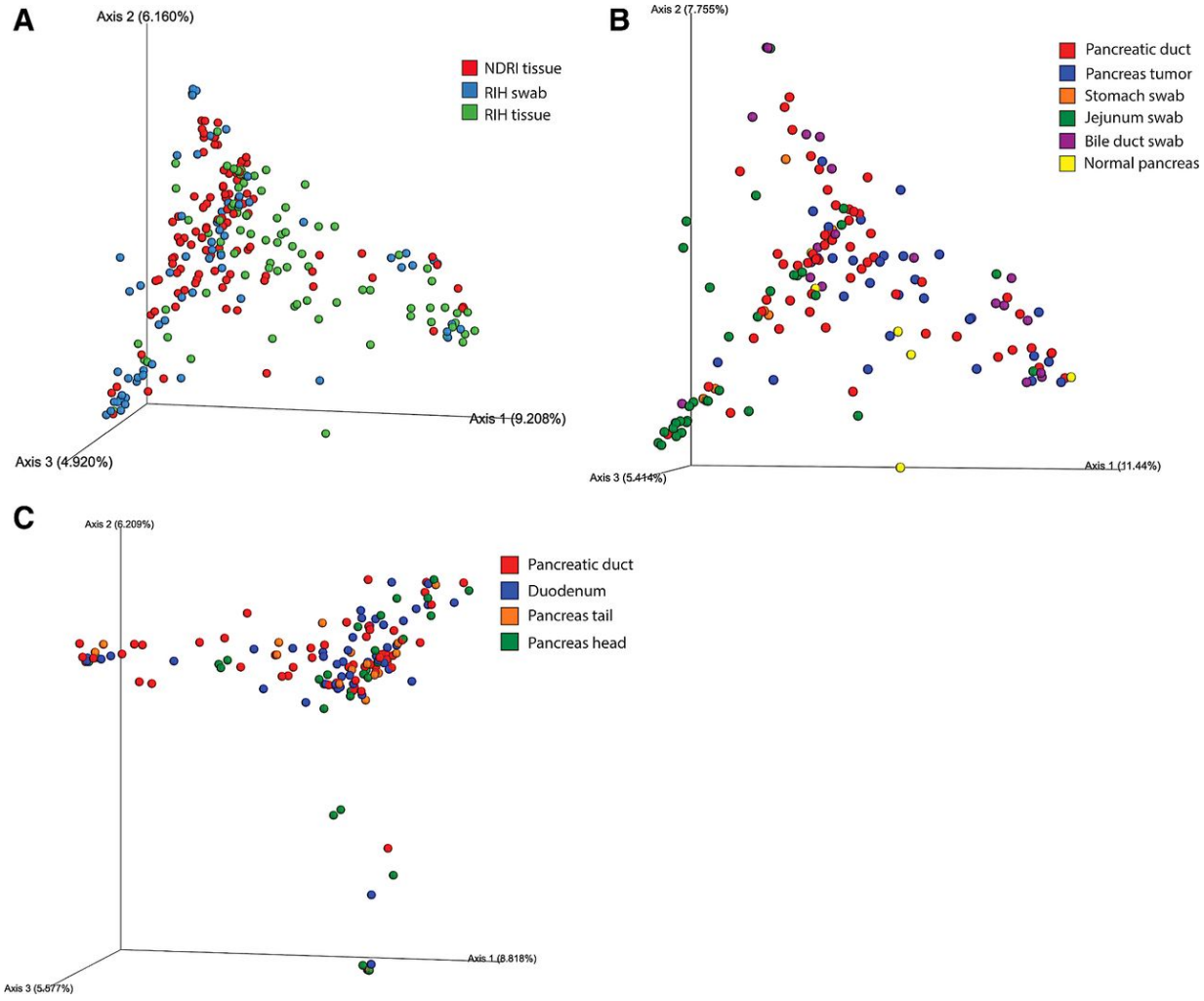


Figure 2.5: PCoA plots showing the relatedness of microbial communities among samples from RIH subjects and NDRI donors using the Bray-Curtis dissimilarity index. Individual datasets are colored according to their (A) RIH and NDRI sample type, (B) RIH anatomical site, and (C) NDRI anatomical site.

likely to be present in cancer tissue samples ($P1 = 0.58$, $p < 0.0001$), and mean relative abundance was higher in non-cancer subjects ($\mu = 0.06$ vs $\mu = 0.02$ in RIH subjects, $p < 0.0001$; Table 2.2). In contrast, a number of bacterial taxa, including *Porphyromonas*, were present in higher mean relative abundance in cancer subjects than non-cancer subjects (Table 2.2, and species level data presented in Supplemental Table A.2). Oral bacteria *Fusobacterium spp*. and *Prevotella spp*. had higher mean relative abundance in cancer subjects than non-cancer subjects (p-values < 0.0001 according to Wald tests for μ). Although these two bacteria do not appear in Table 2.2 because they were not significant according to the joint permutation (based test for prevalence and mean relative abundance at the genus-level), a number of *Fusobacterium* species, e.g., *Fusobacterium nucleatum _subsp._vincentii*, were much more prevalent in RIH samples and are significant in the species-level models (Supplemental Table A.2).

Table 2.3 presents the bacterial taxa (at genus level) for which statistically significant associations remained after multiple comparison correction (at $p < 0.10$) when comparing bacterial taxa in tumor tissue (RIH) by ICD code to those identified in normal pancreatic head tissue from NDRI subjects (labeled as "controls" in Table 2.3; given that the bacterial profiles were highly similar by subject, we included only pancreatic head tissue for this analysis). In the marginal models (prior to adjusting for other covariates), a total of 16 bacterial genera were identified as being significantly associated with disease status prior to correction for multiple comparisons (Supplemental Table A.3); a number of these taxa have representative strains in the Human Oral Microbiome Database (http://www.homd.org) (e.g., *Fusobacterium*, *Capnocytophaga*, *Prevotella*, *Porphyromonas*, *Parvimonas*, *Selenomonas*, and *Haemophilus*). Mean relative abundances for some of these taxa (namely, *Capnocytophaga*, *Prevotella*, *Selenomonas*) were higher in samples coming from subjects diagnosed with pancreatic cancer (ICD C25) compared to NDRI samples. The model with *Porphyromonas* had the strongest association overall ($p = 4.5 \times 10^{-7}$); the relative mean abundance for periampullary cancer tissue samples was substantially higher than that of NDRI samples ($p = 5.8 \times 10^{-19}$), as were the IPMNs (K86.2) samples ($p = 3.6 \times 10^{-7}$). The associations with *Porphyromonas* remained elevated in multiple regression models (Table 2.3).

Table 2.2: Results from multivariable zero-inflated β regression models comparing bacteria presence/absence and relative abundance in tissue and swab samples from NDRI and RIH subjects[a]

			Estimated mean relative abundance (μ)[b]			Estimated proportion of presence (P1)			Global perm test[a]		
Genus	Total read counts	Nonzero samples	RIH	NDRI	Wald P-value	RIH	NDRI	Wald P-value	p-value	AIC difference	P-adjusted[c]
Lactobacillus	1,075,844	154	0.0209	0.0640	<0.0001	0.5270	0.9846	<0.0001	<0.002	35.22	<0.02
Pseudomonas	137,618	90	0.0077	0.0026	<0.0001	0.5523	0.2480	0.0001	<0.002	13.55	<0.02
Parvimonas	45,705	73	0.0091	0.0048	0.0157	0.5251	0.1813	<0.0001	<0.002	10.41	<0.02
Acinetobacter	13915	152	0.0083	0.0040	<0.0001	0.7606	0.5069	0.0008	<0.002	10.12	<0.02
Ralstonia	532	61	0.0001	0.0002	<0.0001	0.1687	0.4098	0.0010	<0.002	8.24	<0.02
Kluyvera	36,883	54	0.0097	0.0045	0.0353	0.2536	0.0361	<0.0001	<0.002	6.94	<0.02
Bilophila	102,226	60	0.0070	0.0013	<0.0001	0.2000	0.0197	0.0001	<0.002	2.95	<0.02
Gemella	76,895	133	0.0193	0.0064	<0.0001	0.7648	0.6134	0.0249	0.004	5.21	0.03
Slackia	34,060	82	0.0084	0.0266	<0.0001	0.2400	0.3958	0.0223	0.008	4.72	0.05
Lachnoanaerobaculum	19,696	92	0.0037	0.0029	0.2536	0.6163	0.2944	0.0008	0.012	0.80	0.07
Solobacterium	16,204	67	0.0069	0.0019	<0.0001	0.5132	0.2351	0.0041	0.014	0.02	0.07
Blautia	156,957	92	0.0030	0.0027	0.4899	0.3081	0.4915	0.0193	0.020	1.62	0.09
Porphyromonas	20,741	86	0.0046	0.0028	0.0192	0.3639	0.5223	0.0465	0.022	0.86	0.10
Anaerococcus	49,115	63	0.0053	0.0069	0.2285	0.1708	0.3281	0.0164	0.026	0.56	0.11
Selenomonas	2,407	73	0.0002	0.0002	0.1128	0.4335	0.2425	0.0205	0.038	0.86	0.15
Staphylococcus	303,413	196	0.0105	0.0175	0.0001	0.8357	0.9146	0.1077	0.042	5.99	0.15
Megasphaera	21,221	69	0.0028	0.0017	0.0266	0.4267	0.2150	0.0134	0.046	-1.45	0.15
Actinomyces	34,042	153	0.0064	0.0034	<0.0001	0.8013	0.7065	0.1642	0.052	-1.62	0.15
Prevotella	222,237	179	0.0240	0.0125	<0.0001	0.8790	0.9046	0.5502	0.052	-1.70	0.15
Bifidobacterium	12,181	88	0.0021	0.0016	0.1358	0.3098	0.1140	0.0073	0.052	-2.33	0.15
Abiotrophia	1,086	43	0.0008	0.0001	<0.0001	0.1502	0.0666	0.0421	0.054	-1.42	0.15
Rothia	227,122	173	0.0274	0.0166	0.0022	0.9458	0.8861	0.0537	0.080	-0.24	0.22

NOTE: Taxonic classification:
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae_[XIII];g__Parvimonas
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ralstonia
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Kluyvera
k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Bilophila
k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Gemellaceae;g__Gemella
k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Eggerthellales;f__Eggerthellaceae;g__Slackia
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Lachnoanaerobaculum
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Solobacterium
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia
k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Porphyromonadaceae;g__Porphyromonas
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae_[XIII];g__Anaerococcus
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Selenomonas
k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Aerococcaceae;g__Abiotrophia
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Megasphaera
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces
k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Prevotellaceae;g__Prevotella
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Bifidobacterium
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia

a All models are adjusted for age, sex, BMI, and log library size. Only bacteria (at genus-level) associated with source of samples at $P \leq 0.10$ before correcting for multiple comparisons are shown. Permutation testing accounts for within subject correlation via random intercept.

b Among nonzero samples.

c Adjusted for multiple testing.

Table 2.3: Results from multivariable zero-inflated β regression models comparing bacteria presence/absence and relative abundance in tissue and swab samples from NDRI and RIH subjects[a]

				Estimated mean relative abundance (μ)[b]				Estimated Proportion of Presence ($P1$)					
Genus	Total read count	Nonzero samples	Control (N=29)	C24 (N=7)	C25 (N=16)	K86.2 (N=6)	Control (N=29)	C24 (N=7)	C25 (N=16)	K86.2 (N=6)	p-value	AIC difference	P-adjusted[c]
Simonsiella	231	15	0.0001	0.0001	0.0018	0.0112	0.1640	0.5559	0.5140	0.1574	<0.0001	35.41	<0.0001
Helicobacter	1,475	7	0.4670	<0.0001	0.0001	0.0001	0.0079	0.6287	0.0005	0.0039	<0.0001	35.38	<0.0001
Porphyromonas	7,008	15	0.0005	0.0512	0.0022	0.0087	0.3886	0.1261	0.0351	0.3175	<0.0001	29.01	<0.0001
Capnocytophaga	316	12	<0.0001	<0.0001	0.0001	0.0017	0.1375	0.2673	0.0865	0.3766	<0.0001	28.87	<0.0001
Ralstonia	153	15	0.0008	0.0061	0.0452	0.0010	0.2955	0.0008	0.6080	0.6144	<0.0001	17.80	0.0026
Bilophila	10,685	13	0.0002	0.0663	0.0139	0.0002	0.0696	0.0547	0.0099	0.1653	<0.0001	17.76	0.0027
Pseudomonas	64,128	31	0.0131	0.0231	0.0469	0.5237	0.2139	0.8512	0.6192	0.6193	0.0001	15.35	0.0076
Acinetobacter	7,916	43	0.0192	0.0615	0.1637	0.1208	0.4407	1.0000	0.7870	0.7825	0.0002	13.81	0.0147
Gemella	7,769	24	0.0028	0.0113	0.0113	0.0071	0.2778	0.5284	0.8466	1.0000	0.0006	11.83	0.0343
Enterococcus	28,254	29	0.0230	0.0176	0.0067	0.0173	0.4277	1.0000	<0.0001	0.4857	0.0007	11.35	0.0419
Propionibacterium	19	10	<0.0001	<0.0001	0.0001	<0.0001	0.1589	0.1454	0.0105	0.0597	0.0011	10.29	0.0655
Peptoclostridium	4,137	14	0.0031	0.0191	0.1199	0.0088	0.2598	0.0003	0.2297	0.5380	0.0017	9.19	0.1034
Solobacterium	1,402	10	0.0001	0.0002	0.0006	0.0003	0.2789	0.8483	0.1583	0.1326	0.0038	7.20	0.2340
Salmonella	37	7	0.0003	0.6282	0.0001	0.0015	0.0336	0.0159	0.0080	0.1074	0.0089	5.09	0.5455
Lactobacillus	251,585	40	0.1343	0.0717	0.1408	0.0891	0.9565	0.6933	0.4354	1.0000	0.0146	3.85	0.8897
Enterobacter	64,118	30	0.0374	0.0323	0.0303	0.0140	0.4467	0.6746	<0.0001	0.2106	0.0177	3.35	1.00
Lactococcus	3,592	17	0.0018	0.0818	0.8557	0.0009	0.2025	0.0684	0.0855	0.2273	0.0194	3.12	1.00
Clostridium	100,517	29	0.0643	0.0459	0.0416	0.0697	0.4600	0.7216	0.0470	0.5243	0.0267	2.27	1.00
Bacteroides	153,955	34	0.0186	0.0171	0.0818	0.0470	0.3719	0.8799	0.7322	0.7174	0.0381	1.33	1.00
Raoultella	31,688	9	0.0593	0.1574	0.0041	0.0081	0.0010	0.0040	<0.0001	<0.0001	0.0523	0.47	1.00

NOTE: Taxonic classification:
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Simonsiella
k__Bacteria;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales;f__Helicobacteraceae;g__Helicobacter
k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Porphyromonadaceae;g__Porphyromonas
k__Bacteria;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ralstonia
k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Bilophila
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter
k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Gemellaceae;g__Gemella
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__Peptoclostridium
k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Solobacterium
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Salmonella
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Enterobacter
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium
k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Raoultella

a All models are adjusted for age, sex, BMI, and sequencing run. Only bacteria (at genus-level) associated with ICD code (overall) at $P \leq 0.05$ prior to correcting for multiple comparisons are shown. Because of missing BMI on two individuals, numbers for the fully adjusted models were based on 58 tissue samples. Marginal models with all samples are shown in Supplementary Table A.3.

b Among nonzero samples.

c Adjusted for multiple testing.

The multivariable regression models for the pancreatic tissue samples identified bacterial taxa (at the genus-level) that had not been significant in the marginal regression models, including *Simonsiella*, *Helicobacter*, and *Bilophia* (Table 2.3 vs Supplemental Table A.3). *Helicobacter* was commonly identified in periampullary pancreatic tumors (C24) but at very low levels; in contrast, *Helicobacter* was infrequently identified in the NDRI samples, but was a dominant genus when present (relative mean abundance 47%; Table 2.3).

We further examined the RIH pancreatic tumor tissue samples without the NDRI samples given the difference in source of tissue and to account for clinical factors such as prior chemotherapy. *Porphyromonas* were also strongly associated with ICD code in both marginal (Supplemental Table A.4) and multiple regression models suggesting clinical covariates were not confounding the main findings for these bacteria.

To test whether the associations would be similar using pancreatic duct tissue samples (vs tumor tissue), we repeated the analysis using RIH and NDRI samples obtained from the pancreatic ducts. The associations for *Porphyromonas* remained detectable and statistically significant in these analyses ($p = 1.53 \times 10^{-11}$).

Using tissue samples obtained from the duodenum, we compared relative abundance of bacterial taxa in NDRI and RIH subjects to examine whether any bacteria from the pancreatic tissue analyses were also noticeably different in the duodenum samples. Of the significant associations noted in the pancreatic tissues, *Selenomonas* was also elevated in the duodenum tissue of pancreatic cancer subjects compared to duodenum tissue from NDRI subjects ($p = 3.9 \times 10^{-12}$). A weak association was also observed for *Gemella* for the duodenum samples, consistent with an overall elevated mean relative abundance in the RIH samples compared to the NDRI samples (Table 2.2); other associations were either not significant or not consistent in direction of differences.

We only had one pancreatic duct stent to examine microbial community; the bacterial taxa from this stent were characterized as the members of the genera *Klebsiella* and *Enterobacter*.

2.6 Discussion

Using pancreatic and duodenum tissue samples from subjects with pancreatic cysts or pancreatic cancer, and comparing them to pancreatic tissue samples obtained from donors who died of non-cancer causes, we were able to demonstrate that pancreatic tissue contains a number of different bacterial taxa, including taxa that are known to inhabit the oral cavity. Our findings provide evidence that the pancreas is not a sterile organ and that there is substantial between-person variability in relative abundance of bacterial taxa at the genera level in the pancreas, but we also observed marked within-person stability across site (Figures 2.1 and 2.2); bacterial composition at different sites in the pancreas (i.e., duct, head and tail) as well as the duodenum were highly similar in the same individuals. Finally, we noted lower presence and relative abundance of *Lactobacillus* in cancer subjects compared to non-cancer subjects, and a significant increase in the mean relative abundance of periodontal-related pathogens in the tissue of pancreatic subjects when compared to non-cancer subjects.

Dissemination of oral bacteria to different parts of the body has been well-reported, and oral bacteria have been linked to a number of chronic diseases, including cardiovascular diseases (LaMonte et al., 2017; Gibson et al., 2006). *Fusobacterium nucleatum* has been associated with colon cancer in a number of cross-sectional studies (Kostic et al., 2011; Castellarin et al., 2011). Mouse models of colorectal cancer provide some support for a causal link (Bullman et al., 2017; Kostic et al., 2013), demonstrating how this bacterium has the ability to initiate recruitment of tumor-infiltrating immune cells. Moreover, a recent study demonstrated similar microbiome profiles in primary colon cancer tumors and liver metastases from the same individuals (resected at a later time point), especially for *Fusobacterium* positive tumors (Bullman et al., 2017), suggesting stability in the microbiome as the tumor progresses and metastasizes. Given the findings from this study, where multiple tissue specimens were examined in the same subjects, it may also be plausible that each individual has a unique microbiome profile that exists in different gastrointestinal tissue and that certain profiles increase cancer susceptibility by impacting the immune environment to allow for tumor promotion and growth. Bacterial taxa found in this study were highly consistent

with those reported in a microbiome study on colon cancer; enriched bacterial taxa associated with *Fusobacterium nucleatum* positive tumors were similar to those we identified in this study (e.g. *Bacteroides*, *Prevotella*, *Selenomonas*, and *Leptotrichia*) (Bullman et al., 2017).

Presence of *Lactobacillus spp*. was significantly reduced in both periampullary and PDAC cancers compared to non-cancer patients (including those with pancreatic cysts). Certain strains of this bacterium have been identified as playing a key role in mediating anti-inflammatory pathways in calorie-restricted mice (Pan et al., 2018). Further research on the role of these bacteria in pancreatic cancer should be conducted.

Previous studies have reported associations between periodontal disease pathogens and pancreatic cancer risk, especially *Porphyromonas gingivalis* (Michaud et al., 2012a; Fan et al., 2016). Periodontal disease is an inflammatory disease of the gums that can, in advanced conditions of periodontitis, result in systemic inflammation. In this study, we observed significantly higher mean relative abundance levels (at the genus-level) for two bacterial taxa previously associated with periodontitis in pancreatic tissue, including *Porphyromonas* and *Selenomonas* (Faveri et al., 2009; Stingu et al., 2012; Liu et al., 2012; Gonçalves et al., 2012); however, only *Porphyromonas* remained statistically significant after adjusting for age, sex, BMI and library size. *Porphyromonas* was also elevated in the pancreatic duct tissue of periampullary pancreatic cancers, but no statistically significant associations were noted for the other oral bacterial taxa. Whether *Porphyromonas* play a role in pancreatic carcinogenesis will need to be further examined in other studies and confirmed in animal models. Proposed mechanisms for carcinogenesis include the ability of certain bacteria to induce a pro-inflammatory response in the tumor microenvironment (Kostic et al., 2013); inhibit the immune response targeted at eliminating tumor cells (Gur et al., 2015); and modulate key cellular pathways associated with cell division (Rubinstein et al., 2013).

A similar study using swab specimens from the pancreas, bile and jejunum, was conducted on subjects with pancreatic cancer undergoing pancreaticoduodenectomy (Rogers et al., 2017). In that study, many bacterial taxa were present in fluids obtained from the pancreatic ducts and the common bile duct, including *Prevotella*, *Haemophilus*, *Aggregatibacter*, and *Fusobacterium*

(Rogers et al., 2017). Consistent with our findings, microbial communities in the pancreas, bile and jejunum fluids were similar within individuals (Rogers et al., 2017). Mean relative abundance for the bacterial genus *Klebsiella* was high in the samples from pancreatic cancer subjects in that study (Rogers et al., 2017); in our study, we found *Klebsiella* to be one of two taxa on a swab taken from the stent itself. Placement of stent prior to surgery may impact the type of bacteria present in the pancreas, as observed in our study. In a separate study, metagenomics was conducted on freshly frozen duodenum samples from 5 normal and 5 obese individuals; *Streptococcus* $(30-32\%)$ and Actinomyces $(12-17\%)$ were the most common bacterial taxa identified in those samples, and relatively higher counts of *Gemella* were also identified in all 10 subjects (Angelakis et al., 2015). *Porphyromonas* were not identified in the duodenal samples (Angelakis et al., 2015).

In a recent study examining tumor resistance to the drug gemcitabine, bacteria were found in tumor tissues of 65 PDAC patients (out of 113), and 51.7% of bacterial taxa belonged to the class Gammaproteobacteria (Geller et al., 2017), which is highly consistent with our findings (Figure 2.1). Similar to our study, there was large inter-individual variability in relative abundance of bacteria in each tumor, but in contrast to our study, only 3 out of 20 organ donors were found to be positive for bacteria, and no normal tissue samples were included from the same patients (Geller et al., 2017). In addition, a high number of reads for *Porphyromonas* was found in one (out of 65) pancreatic cancer tissue specimens (mean relative abundance of 0.123; Supplementary Material (Geller et al., 2017)). In our study, read counts for *Porphyromonas spp*. were also extremely high in two RIH subject. In a separate study, 408 genera of bacteria were identified in pancreatic cyst fluids obtained from patients through endoscopy (Li et al., 2017); many of the taxa found in pancreatic cysts were similar to those in tissue from our study, including the presence of *Fusobacterium*. Furthermore, *Porphyromonas* was present in 33% of fluid samples and relative abundances for those taxa were similar to those in our study (non-zero cysts mean relative abundance: 0.00178, range 0.0001-0.004) (Li et al., 2017).

In a recent study, *Bifidobacterium spp*. was found to increase in abundance in the feces of mice with Kras mutations (genetically modified to increase pancreatic cancer) as disease progressed,

compared to wildtype mice (Pushalkar et al., 2018). Furthermore, gut repopulation of the germ-free (Kras) mouse with *Bifidobacterium pseudolongum* increased T-cell infiltration and tumor growth (Pushalkar et al., 2018). Similarly, we also noted a higher prevalence for the genus *Bifidobacterium* in cancer subjects compared to non-cancer subjects (Table 2.2).

Several studies have looked at the involvement of bacteria in biliary and pancreatic diseases and have observed a high number of bacterial taxa present in the calcified pancreatic duct epithelium and in pancreatic abscess (Swidsinski, 2005; Schmid et al., 1999; Brook & Frazier, 1996; Hill et al., 1983; chiang Tsui et al., 2009). Anaerobic bacterial taxa have been found at a variable rate in pancreatitis; the results depend on the process for bacterial identification (Swidsinski, 2005; Schmid et al., 1999; Brook & Frazier, 1996). Previous studies have also reported the presence of bacteria in bile (Wu et al., 2013; Ye et al., 2016). In a study of 6 subjects with gallstones, 16S rRNA gene sequencing identified high relative abundances of *Escherichia*, *Klebsiella* and *Pyramidobacter* in the bile, and the bacterial profile of the bile was very similar to the duodenum in the same subjects (Ye et al., 2016). *Pyramidobacter* species was originally isolated from the oral cavity (Downes et al., 2009) and was also found in our study samples, but at low levels ($< 20\%$ of all samples).

Several bacterial taxa we observed with elevated relative mean abundance in RIH samples have been previously identified in immunocompromised patients and are largely believed to be opportunistic pathogens, including *Acinetobacter* (Bergogne-Berezin & Towner, 1996) and *Kluyvera* (Sarria et al., 2001). The genus *Gemella*, which was found at higher relative abundance in pancreatic cancer subjects when compared to NDRI samples, has been previously associated with a number of infections, including endocarditis, soft-tissue abscesses, empyema, bloodstream infection, and bone infections (Mosquera et al., 2000; Scola & Raoult, 1998; García-Lechuz et al., 2002; Fangous et al., 2016). Because our analysis was based on a cross-sectional study design, we expected to identify bacteria that were present as a result of opportunistic nosocomial infections given that the majority of RIH subjects were likely immunocompromised from their cancer. However, our results show that even normal pancreatic tissue harbors a microbial community.

The strength of this study was the collection of specimens specifically for the purpose of microbiome analysis, with precautions made to reduce contamination during collection and processing of samples. Moreover, multiple types of samples were collected on each patient at RIH, including obtaining tissue or swabs from multiple sites, to allow for inter vs. intra-individual differences at different sites. Finally, the multivariable regression analyses was conducted to adjust for potential confounding by known pancreatic cancer risk factors, including BMI and smoking, as well as other factors that may cause bias, including pre-OP EUS and prior chemotherapy.

The major limitation of this analysis was the small number of subjects with pancreatic cysts and pancreatic cancer; despite recruiting 77 subjects, not all subjects had tissue resections during surgery (as more advanced pancreatic cancer patients are often not operable). We did not have sufficient power to examine in great detail the differences in bacterial composition between different pancreatic cancer subtypes, including IPMNs; however, we were the first to include ICD 24 tumors and to explore differences with ICD 25 tumors. Moreover, cancer versus non-cancer comparisons of bacterial presence/absence and relative abundances were based on subjects spread across two different data sources (i.e., RIH and NDRI). Differences in microbiota between these two sources may have been due to differences in collection methods and collection times; DNA was extracted from frozen tissue using the same protocol and methods, but tissue samples were either collected during surgery (RIH) or from organs that were rapidly frozen after death (NDRI samples had a mean time of 13 hours to processing of samples). Consequently, it is possible that the identified genera (and overall differences in bacterial taxonomy) merely reflect study-specific differences, rather than real cancer-specific differences.

In this culture-independent study, we detected many bacterial taxa in pancreatic tissue from cancer subjects as well as non-cancer subjects. Furthermore, the bacterial profiles in the pancreas were more similar within individuals across different sites of the pancreas (i.e., head, tail, ducts) and duodenum than between individuals at each site. Bacterial taxa known to inhabit the oral cavity were common in the pancreas microbiome and several periodontal pathogens were also identified in pancreatic tissue samples. Further research is needed to address if and how bacteria may be

related to pancreatic carcinogenesis or disease progression.

Chapter 3

A Bayesian framework for identifying consistent patterns of microbial abundance between body sites

This chapter has previously been published as an open access article and is reprinted here with minor adaptations. Meier, R., Thompson, J.A., Chung, M., Zhao, N., Kelsey, K.T., Michaud, D.S., Koestler D.C. (2019). A Bayesian framework for identifying consistent patterns of microbial abundance between body sites. *Stat. Appl. Genet. Mol. Biol.* 18(6).

3.1 Statement of Contributions

In this project, I, Richard Meier, developed the methodology, conducted the simulation studies and statistical analyses, and wrote the manuscript. Contributions of other authors are listed below.

DSM managed the acquisition and processing of the reference data sets. JAT provided advice and guidance on the statistical methodology and edited the manuscript. MC, NZ and KTK assisted in the writing of the manuscript and interpretation of the study findings. DSM is the principal investigator of the pancreatic microbiome study and assisted in the interpretation and drafting of the manuscript. DCK helped conception of the methodology, supervised the implementation, and edited the manuscript. All authors read and approved the final version of the manuscript.

3.2 Abstract

Recent studies have found that the microbiome in both gut and mouth are associated with diseases of the gut, including cancer. If resident microbes could be found to exhibit consistent patterns between the mouth and gut, disease status could potentially be assessed non-invasively through profiling of oral samples. Currently, there exists no generally applicable method to test for such associations. Here we present a Bayesian framework to identify microbes that exhibit consistent patterns between body sites, with respect to a phenotypic variable. For a given operational taxonomic

unit (OTU), a Bayesian regression model is used to obtain Markov-Chain Monte Carlo estimates of abundance among strata, calculate a correlation statistic, and conduct a formal test based on its posterior distribution. Extensive simulation studies demonstrate overall viability of the approach, and provide information on what factors affect its performance. Applying our method to a dataset containing oral and gut microbiome samples from 77 pancreatic cancer patients revealed several OTUs exhibiting consistent patterns between gut and mouth with respect to disease subtype. Our method is well powered for modest sample sizes and moderate strength of association and can be flexibly extended to other research settings using any currently established Bayesian analysis programs.

3.3 Introduction

Microbial communities inhabit virtually every part of the human body and can differ across individuals. Even within the same individual, microbial communities often change with anatomical location (Faith et al., 2013). In this context, it is not surprising that the human microbiome plays an important role in a wide range of diseases, including even life threatening conditions such as cancers. In their review, Goodman & Gardner (2018) summarize several compelling examples, such as increased *Fusobacterium* species associating with tumors in colon and *Helicobacter pylori* inducing lymphoma and gastric cancer. More recently, bacteria have been identified in pancreatic tissue in cancer patients (del Castillo et al., 2019) and have been shown to play a role in carcinogenesis in the pancreas (Pushalkar et al., 2018). Additional studies have also reported evidence that certain oral bacteria and periodontal disease associate with an increased risk in pancreatic cancer (Michaud et al., 2012b; Fan et al., 2016). Finally, it has been shown that *Fusobacterium nucleatum*, a common oral bacterium, produces a protein that allows itself and other bacteria to travel through the endothelium in the mouth and into the blood stream, allowing them to migrate to other body sites (Fardini et al., 2011). Despite the empirical evidence, little is understood about how these associations originate and no confirmatory study has conclusively established their biological mechanism. This motivates the question of whether microbes exist for which changes in

abundance (mean relative abundance or rate of presence) with respect to disease status in the oral cavity correspond to changes in abundance in gut samples. In other words, are there microbes for which fluctuations in their abundance are preserved across disease status between mouth and gut? Identification of such species, exhibiting **pa**irwise **st**ratified **a**ssociation (PASTA) between two body sites, may allow further insight into mechanisms and the biology behind a disease. Furthermore, it may also provide new opportunities for treatment or detection and even potentially enable a researcher or medical professional to learn about the disease in the gut by monitoring oral samples. Considering that gut samples can only be acquired through invasive surgical procedures, PASTA microbes could constitute invaluable clinical markers.

Data arising from 16S rRNA sequencing for assessing the microbiome takes the form of compositional count tables. The term operational taxonomic unit (OTU) can be understood as a group of closely related microbes on a given taxonomic level, for example: phylum, genus, or species. For a given experiment, in which abundance of microbes is quantified in a series of biological samples, each cell in row i and column j of such a table represents how often a species or OTU i was observed in sample j (Table 3.1) Unfortunately, these data are intricate with total column counts (sequencing depth and microbial yield) differing between samples, high frequency of zero values (i.e. sparsity), and the constant sum constraint problem that can create spurious associations when few rows dominate the majority of counts (Tsilimigras & Fodor, 2016; Gloor et al., 2017).

Due to its complexity, many different modeling strategies have been proposed for the analysis of microbial 16S rRNA abundance data. When investigating an individual microbe (or a specific group of microbes), current strategies predominantly aim to understand the relationship between abundance and selected phenotypes. Three major parametric approaches are employed by most researchers: discrete data models such as Zero-inflated Poisson or Zero-inflated Negative Binomial regression (Xia et al., 2018; Zhang et al., 2017); log-ratio Aitchison models that explicitly address the constant sum constraint by treating the ratio of abundance counts between two taxonomic units as the response (Shi et al., 2016; Tsilimigras & Fodor, 2016; Gloor et al., 2017); and lastly, relative abundance models that transform counts into sample proportions and fit semi-continuous

models to the data such as Zero-inflated Beta regression (ZIBR) (Xia et al., 2018; Peng et al., 2016; Chen & Li, 2016). Each approach can present specific advantages and limitations, where the most suitable model will depend on the circumstances of the research study. While log-ratio Aitchison models are mandatory in datasets either measuring high phylogenetic levels with few taxonomic units or exhibiting low community diversity (Tsilimigras & Fodor, 2016), discrete data and relative abundance models are convenient to address sparsity in high diversity settings. To date, neither of these modeling strategies has been utilized to test for PASTA relationships and there presently exists no general testing approach that is applicable regardless of the parametric modeling strategy. Alternatively, non-parametric inter-rater strategies can be employed to test for agreement or association between body-sites. These strategies assume that there are individual raters that are presented with two different scenarios or cases, each of which they have to assign to either a category or numeric value. The methods then ask the question whether individual raters tend to make assignments that agree or associate between the two scenarios. Popular examples are Cohen's kappa (Cohen, 1960; Fleiss, 1971) for categorical responses and Pearson or Spearman correlation for numeric responses (Schober et al., 2018). These methods do not necessarily require knowledge about the distribution of the response and are applicable even if there is strong disagreement or variability between individual raters. However, they do not allow accounting for confounders or other sources of variation, and they require paired samples.

Table 3.1: Hypothetical example of a microbial abundance data table. Rows represent genera, which are groups of closely related microbes and an example of a type of operational taxonomic unit (OTU). For a given sample and OTU, each cell in the table counts how often said OTU was observed in said sample through the 16S rRNA sequencing technique. All counts in this type of table are expected to increase with the total number of observed OTUs in the respective sample. These column totals can be understood as the sample signal intensity and change based on experimental parameters for each sample.

Genus	Sample 1 Count	Sample 2 Count	Sample 3 Count	Sample 4 Count	... ...
Actynomices	0	0	3	5	...
Atopobium	0	27	10	6	...
Fusobacterium	0	14	0	0	...
...	...	...	...	...	...
SAMPLE TOTAL	671	2390	1502	1883	...

Here, we present an approach to test for PASTA that is applicable regardless of the data model and regardless whether all, some, or none of the samples are paired. The question of PASTA relationships with respect to body site is translated into a question of association of population parameters (such as mean relative abundance) between the two body sites. A test is then proposed based on applying a correlation statistic to parameter estimates. Testing and adjusting for paired samples is made convenient by utilizing a Bayesian modeling framework. For the purpose of illustration, this paper will focus on modeling relative abundance via a ZIBR model, though as stated before, the approach is not limited to any particular data model. After establishing the data model and introducing the approach, viability and performance are evaluated via simulation studies and through the analysis of a biological dataset involving microbiome data collected from the gut and specific oral sites in patients with pancreatic cancer and other diseases of the foregut. Finally, strengths, limitations and opportunities for future methodological development are discussed.

3.4 Methods

3.4.1 Experimental Design

A study suitable to answer the previously described research question can be broken down into the following steps. First, an appropriate subject population exhibiting the disease or target phenotype is identified and biological samples from the two body sites of interest are collected. Multiple samples from the same patient within and across body sites are possible, but not necessarily required. Next, sample preparation and 16S rRNA sequencing are performed. This sequencing technique aims to identify and count hypervariable DNA patterns that are specific to microbial species and OTUs, but that do not exist in human DNA; the rationale being that the DNA content of a group of microbes is approximately proportional to their abundance in the sample. So, by counting how often signatures belonging to a specific OTU are observed, we can obtain an estimate of its abundance relative to how many microbes were observed, in total. After OTUs have been counted, our proposed statistical test is performed individually for each OTU, testing the null hypothesis that there is no PASTA relationship for each specific set of considered microbes. This test is performed

by fitting a statistical regression model to the row vector of abundance values corresponding to a target OTU, followed by the calculation of a test statistic T_θ based on the parameter estimates (for example, rate of absence or mean abundance) obtained from said model. This statistic will be small when H_0 is true and large when H_0 is false. An overview of the experimental design for testing the hypothesis of PASTA can be found in (Figure 3.1).

3.4.2 Data Model

In what follows we consider abundance on two taxonomic levels: the genus and the Amplicon Sequence Variant (ASV) level, the latter representing unique biological sequences that were identified from 16S genes (Callahan et al., 2017). In order to make abundance values comparable across samples and bring them to the same scale, raw counts are first transformed into relative abundance values. For a given sample, relative abundance of an OTU refers to the number of times that OTU was observed, scaled by the total number of observed OTUs for that sample. It represents the proportion of times an OTU was observed in a given sample.

Let Y_k denote the relative abundance of a specific OTU for sample k. This response can be modeled as a Zero-inflated Beta distribution with probability density $f_{Y_k}(y|p_k, \omega_k, \phi_k)$. This model assumes that the case $Y_k = 0$ occurs with probability p_k and that given $Y_k > 0$, the response Y_k follows a Beta distribution with mean ω_k and dispersion ϕ_k. For a given OTU and sample, the probability of absence p defines how likely it is to observe no microbe comprising that OTU within said sample. The mean non-zero relative abundance ω represents the mean relative abundance given that microbes comprising the OTU are actually observed. The mean of Y_k, the overall mean relative abundance, is then $E[Y_k] = \mu_k = \omega_k(1 - p_k)$. The probability density function of this distribution can be expressed as follows:

$$f_{Y_k}(y) = \begin{cases} p_k & \text{if } y = 0 \\ (1 - p_k) \cdot \frac{\Gamma(\phi_k)}{\Gamma(\omega_k \cdot \phi_k)\Gamma((1-\omega_k)\cdot\phi_k)} y^{\omega_k \cdot \phi_k - 1}(1-y)^{(1-\omega_k)\cdot\phi_k - 1} & \text{if } y > 0 \end{cases} \tag{3.1}$$

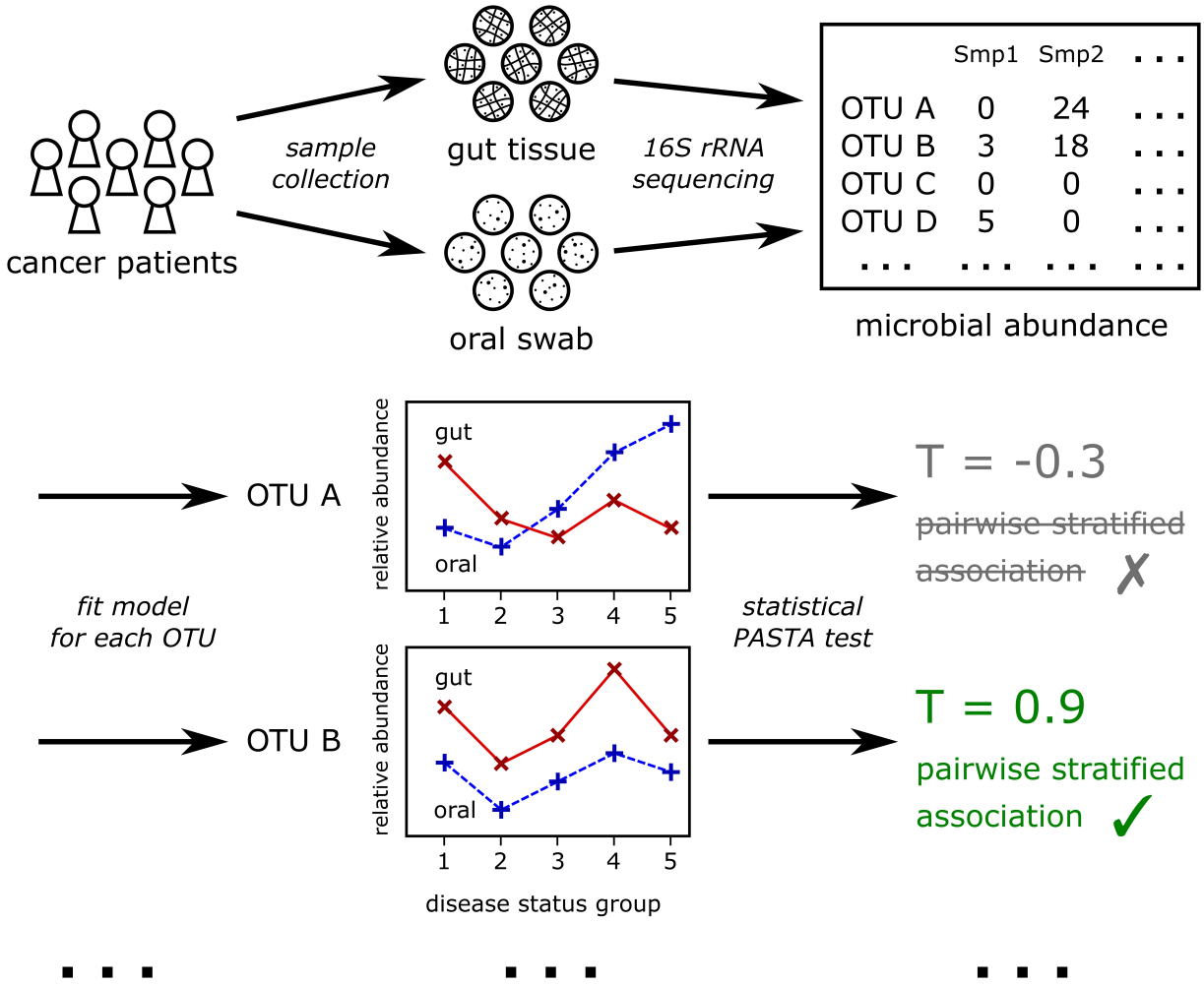


Figure 3.1: Overview of the experimental setup to test for pairwise stratified association (PASTA). Oral and gut samples are obtained from cancer patients and 16S rRNA sequencing is performed on each sample. The resulting microbial abundance data is used to fit a statistical regression model to each observed OTU across all samples. Finally, abundance estimates across strata are used to test whether abundance patterns in disease status are preserved between mouth and gut.

Before the statistical PASTA test can be performed, a Bayesian ZIBR model is fit to the data Y_k utilizing the likelihood f_{Y_k} and assuming a common dispersion parameter $\phi_k = \phi$ for all samples. The estimated posterior distributions of ω and p resulting from this model are then subsequently used to conduct the test.

Let ω denote the vector of mean relative abundances for all samples, $\mathbf{p}$ denote the vector of probabilities of absence for all samples, β, δ denote coefficient vectors, $\mathbf{b}, \mathbf{d}$ denote random effect vectors and $\mathbf{Q}, \mathbf{R}, \mathbf{W}, \mathbf{X}$ represent design matrices. The design matrices code how covariates impact the model parameters via the following link functions:

$$logit(\omega) = \beta\mathbf{X} + \mathbf{bR} \quad \text{and} \quad logit(\mathbf{p}) = \delta\mathbf{W} + \mathbf{dQ} \tag{3.2}$$

For our application, the matrices $\mathbf{W}$ and $\mathbf{X}$ are used to model the strata of body site and disease status, but can additionally be used to adjust for other, fixed covariates (e.g., subject age, gender, smoking status, and/or other potential confounders or sources of variation). On the other hand, the optional inclusion of $\mathbf{Q}$ and $\mathbf{R}$ permits one to account for correlation structures, such as within-subject correlation when multiple samples are collected from the same patient. If, for example, the probability of absence of a given OTU in sample k from subject j_k is assumed to be impacted by disease status g_{j_k}, body site s_k, age A_{j_k} and within-subject correlation, we would formulate $\mathbf{W}$, $\mathbf{Q}$ and our model for the probability of absence as follows:

$$logit(p_k) = \delta_{1,g_{j_k},s_k} + \delta_2 A_{j_k} + d_{1,j_k} \tag{3.3}$$

Here, δ_{1,g_{j_k},s_k} captures the effect of disease status g_{j_k} and body site s_k on the probability of absence, δ_2 represents the effect of age on the probability of absence, and d_{1,j_k} is a subject specific random intercept. Analogously, if mean relative abundance is believed to be impacted by the same covariates in the same way, except that age was believed to have no effect, we would formulate $\mathbf{X}$,

R and our model for non-zero relative abundance such that:

$$logit(\omega_k) = \beta_{1,g_k,s_k} + b_{1,k} \tag{3.4}$$

Here, β_{1,g_{j_k},s_k} captures the effect of disease status g_{j_k} and body site s_k on the mean non-zero relative abundance and b_{1,j_k} is a subject specific random intercept. The specific models used in our analysis of the pancreatic cancer dataset are presented in the Results Section are further discussed in Section 3.4.7.

Let $t, i \in \mathbb{N}$ denote placeholder indices for any potential coefficient specified in the above ZIBR model. Then all posterior distributions were estimated, utilizing the following independent prior distributions:

$$\pi(\beta_t), \pi(\delta_t) \sim N(0, 100);\ \pi(b_{t,i}) \sim N(0, \zeta_t);\ \pi(d_{t,i}) \sim N(0, \xi_t); \tag{3.5}$$

$$\pi(\sqrt{\phi}) \sim Unif(1, 100);\ \pi(\zeta_t^{-1}), \pi(\xi_t^{-1}) \sim Gam(0.01, 0.01) \tag{3.6}$$

Priors were chosen to be weakly informative, with the exception of $\sqrt{\phi}$ being restricted to values larger than or equal to 1 in an attempt to stabilize estimation of means. Under these priors and for some integer vectors $\mathbf{T}, \mathbf{I_1}, \mathbf{I_2}$ the posterior distribution of parameters will then satisfy the following:

$$\begin{aligned}\pi(\beta, \delta, \mathbf{b}, \mathbf{d}, \zeta, \xi | \mathbf{Y}) \propto & \prod_{t_1}^{T_1} \pi(\beta_{t_1}) \cdot \prod_{t_2}^{T_2} \pi(\delta_{t_2}) \cdot \prod_{t_3}^{T_3} \left(\pi(\zeta_{t_3}) \prod_{i_{t_3}}^{I_{1,t_3}} \pi(b_{t_3,i_{t_3}} | \zeta_{t_3}) \right) \cdot \\ & \prod_{t_4}^{T_4} \left(\pi(\xi_{t_4}) \prod_{i_{t_4}}^{I_{2,t_4}} \pi(d_{t_4,i_{t_4}} | \xi_{t_4}) \right) \cdot f(\mathbf{Y} | \beta, \delta, \mathbf{b}, \mathbf{d}, \zeta, \xi)\end{aligned} \tag{3.7}$$

There are generally no analytical solutions for the posterior distributions of the coefficients when random effects are present. Regardless, whether the model structure is a special case that allows for analytical calculation of posterior distributions or whether we employ a more complex model where this is not possible, posterior distributions can be estimated via Markov chain Monte

Carlo (MCMC) methods. Briefly, MCMC procedures allow one to draw arbitrarily large samples from a posterior distribution that will numerically approximate the said distribution, as the number of draws increases. Models were fit via this method in the software OpenBUGS (version 3.2.3 rev 1012) via the R (version 3.4.0) package "R2OpenBUGS" (version 3.2.3.2).

3.4.3 Formal Definition of Pairwise Stratified Association (PASTA)

In order to understand how the Bayesian regression model can be used to conduct the desired hypothesis test, we will first provide a formal definition of PASTA. Let s denote a grouping variable for which two groups are to be compared. For our purposes, this grouping variable represents body sites: $s = 1$ denotes gut and $s = 2$ denotes mouth. Let g denote another grouping variable with three or more distinct categories. This grouping variable will represent different types of disease status, more specifically cancer-subtype. Let θ_{sg} be a population parameter of the response for a given body site s and disease status g. The population parameter represents fundamental properties of the distribution of the response. For the here considered ZIBR model, p, ω and μ are relevant candidates for θ. If PASTA holds for a given OTU, then either p, ω or μ will associate between the two body sites, because they all relate to the magnitude of abundance.

We thus define: The parameter θ exhibits PASTA with respect to s and g if there exists an increasing function $h(x)$ such that $\theta_{1g} = h(\theta_{2g})$ holds for all $g \in \{1, 2, \ldots, G\}$, where $G \geq 3$. Conceptually, this definition says that as we move from one disease status group to another, if θ increases in oral samples, it will also increase in gut samples. Analogously, if θ decreases from one disease status group to another in the mouth, it will also decrease in the gut. A visualization is provided in Figure 3.2.

3.4.4 Testing for PASTA

Let $T(\mathbf{x}, \mathbf{y}) \in [-1, 1]$ denote a correlation statistic between two numerical vectors $\mathbf{x}, \mathbf{y}$; for example, the Pearson or Spearman correlation statistic. Under this definition, $T_\theta = T(\theta_1, \theta_2)$ denotes the correlation statistic calculated for the two parameter vectors corresponding to $s = 1$ (e.g. parame-

body site	disease status			
	$g = 1$	$g = 2$	$g = 3$	$\ldots$
$s = 1$ (gut)	θ_{11}	θ_{12}	θ_{13}	$\ldots$
$s = 2$ (mouth)	$\theta_{21} = h(\theta_{11})$	$\theta_{22} = h(\theta_{12})$	$\theta_{23} = h(\theta_{13})$	$\ldots$

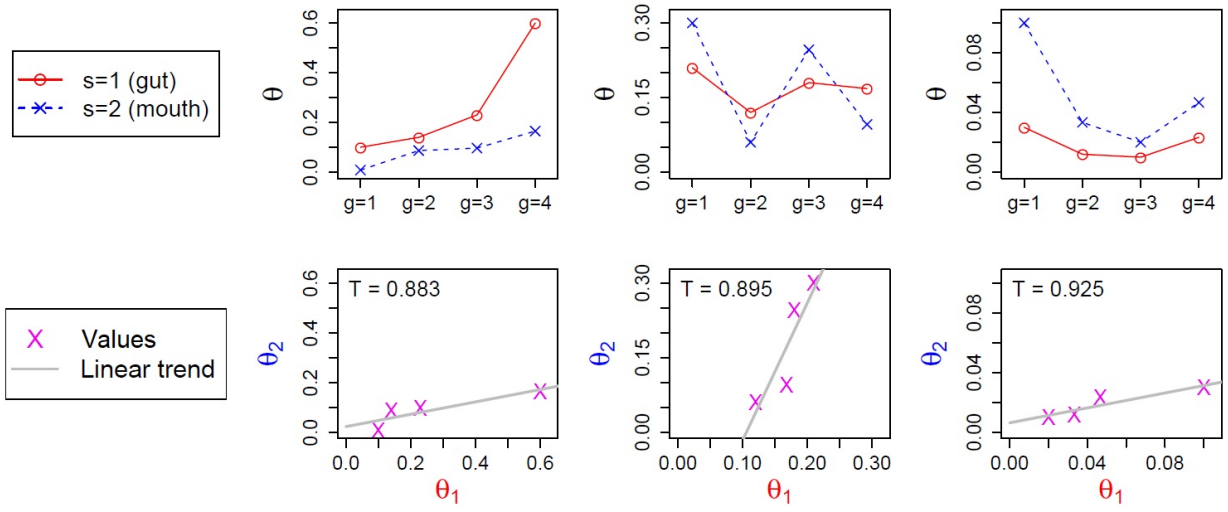


Figure 3.2: Visualization of pairwise stratified association (PASTA). Let θ represent a population parameter of interest, for example the mean relative abundance of a particular OTU. Each column of sub-figures below the table are examples of a PASTA relationship, i.e. of h being an increasing function. The first row plots parameter values of mouth and gut side-by-side and demonstrates that a variety of different scenarios are covered by this definition. In the second row, plotting parameter values of gut against parameter values of mouth reveals their association through a trend. T denotes Pearson correlation values between gut and mouth.

ters for disease status groups in the mouth) and $s = 2$ (e.g. parameters for disease status groups in the gut). Generally, if a PASTA relationship holds between θ_1 and θ_2, this statistic should assume a larger value compared to cases where such a relationship does not hold. This means that we are able to formulate our desired test by rejecting H_0 if T_θ is larger than a specific threshold and fail to reject H_0 if it is less than said threshold.

In summary, assume that $\theta_{1g} = h(\theta_{2g})$ implies $T_\theta > t_c$ for some $-1 < t_c < 1$. The constant t_c represents a meaningful degree of association. For example, a value of $t_c = 0$ would mean that any tangible degree of association is meaningful, where a value of $t_c = 0.5$ would mean that a moderate degree of association is meaningful. This definition is useful because T_θ can score the degree of association without explicitly having to specify the shape of h. Considering the complexity of

the biology underlying the samples, specifying h in advance may not only be hard to justify, but strong deviations of a chosen h from the true h could also result in missing promising associations. Instead, our regression model will allow each stratum (s,g) to have an independent effect on the response, leading to a unique, agnostic posterior distribution of each θ_{sg}. These unique posteriors are then in turn used to calculate the posterior distribution of T_θ and conduct the hypothesis test.

Based on this scoring definition of PASTA, we formulate our hypotheses in the following way:

$H_0 : T_\theta \leq t_c$, i.e. θ_1 and θ_2 do NOT exhibit PASTA

$H_1 : T_\theta > t_c$, i.e. θ_1 and θ_2 DO exhibit PASTA

While deriving analytical solutions of the distribution of $T_\theta|\mathbf{Y} = T(\theta_1|\mathbf{Y}, \theta_2|\mathbf{Y})$ will depend on the data model and may be difficult or even impossible to obtain depending on the modeling scenario, a general testing procedure can still be derived. As described earlier, MCMC methods allow one to conveniently obtain a large sample of posterior draws of each θ_{sg}, even when obtaining analytical solutions of posterior distributions is not possible. Furthermore, plugging the posterior draws of each MCMC iteration into T allows one to obtain posterior draws from T_θ itself. Let α denote the target credibility threshold, H_0 is then rejected if the lower bound $t_{Q\alpha}$ of the one-sided credible interval of $T_\theta|\mathbf{Y}$ exceeds t_c. This is equivalent to rejecting H_0 if the estimated probability of no association exceeds α, i.e. $Pr(T_\theta|\mathbf{Y} \leq t_c) < \alpha$. In detail, the step by step process for testing H_0 is as follows:

1. Specify a likelihood for the response data $\mathbf{Y}$ and prior distributions for the parameters θ

2. Utilize a MCMC sampling scheme to draw a large number of samples from the posterior distributions of the parameters $\theta|\mathbf{Y}$. One draw from the Markov Chain contains a unique draw for each θ_{sg}.

3. Calculate $T_v^* = T(\theta_1^{*v}, \theta_2^{*v})$ where θ^{*v} denotes the v^{th} MCMC draw. Then $\mathbf{T}^*$ is a large sample of the posterior distribution $T_\theta|\mathbf{Y}$.

4. Calculate the $\alpha \cdot 100\%$ sample quantile $t_{Q\alpha}$ of $\mathbf{T}^*$. If the Markov Chain is sufficiently long,

the sample quantiles of $\mathbf{T}^*$ will closely approximate the quantiles of the true posterior distribution. The value $t_{Q\alpha}$ is thus the lower bound of the $(1-\alpha)\cdot 100\%$ one-sided credible interval of $T_\theta|\mathbf{Y}$.

5. Reject H_0 if the lower bound $t_{Q\alpha}$ is larger than t_c.

This process is generally applicable regardless of the data model or the parameter being tested, as long as each θ_{sg} can be estimated without constraining them to a parameter space that implies PASTA.

3.4.5 Pancreatic Cancer Patient Dataset

In order to evaluate validity of the approach in the context of microbiome data, analyses were performed based on a biological 16S rRNA sequencing dataset first published in del Castillo et al. (2019). This dataset contained samples of various gut and oral sites from 77 patients with pancreatic cancer with age range 31 to 86 years. Sequencing was performed utilizing the Illumina MiSeq System and alignments were performed using BLASTN against a reference library combining sequences from HOMD (version 14.5), Greengenes Gold and the NCBI 16S rRNA reference sequence set. OTU counts were obtained utilizing the QIIME (Quantitative Insights Into Microbial Ecology 22) software package version 1.9.1, while the unique Amplicon Sequence Variant (ASV) counts were calculated using the QIIME2 software package release 2018.4. The former was used to obtain taxonomic genus level counts, whereas the latter was used to obtain rarefied ASV level information, calculated based on sequencing data rarefied at a sampling depth of 1200. Both genus level and ASV level counts were considered for analysis. Before fitting statistical models to the data, relative abundance values of less than 0.01 were treated as noise and set to 0. To ensure inference was based on sufficient signal, OTUs and ASVs were only tested if more than 5% of all samples exhibited non-zero values.

The dataset was used to both guide simulation studies (described in the next section) and to deploy models to identify potential microbes that may exhibit a PASTA pattern.

3.4.6 Simulation Studies

Before simulations were performed, an empirical approach was pursued in order to obtain sampling distributions of the parameters p, ω, ϕ that would be representative of biological microbiome data. First, a marginal, unstratified ZIBR model was fit to the pancreatic cancer dataset that assumed all samples of relative abundance for a given OTU originated from the same distribution. These model fits yielded a single estimate of p, ω and ϕ for each OTU. These estimates were then assumed to be representative of or approximate the true distribution of parameters in biological data. In the next step, the estimates were used to obtain smooth probability distributions that parameters could be sampled from during the simulation studies. For both p and ω, individual Beta distribution models were fit to the marginal estimates in order to obtain their smooth sampling distributions. On the other hand, $\log \phi$ was sampled via a Normal distribution through an observed linear relationship between $\log \phi$ and $\log \omega$ that was present on the ASV level and the genus level. More specifically, since our models assumed fixed dispersion among all groups, dispersion was sampled from $(\log \phi | \min_{sg}\{\log \omega_{sg}\}) \sim N(a \min_{sg}\{\log \omega_{sg}\} + b, \sigma^2)$, where a, b, σ^2 differed between genus and ASV level.

After the smooth sampling distributions were obtained, the performance of PASTA tests was evaluated via simulations. Let t denote a target, fixed degree of association, n denote the number of observations in each stratum (s, g) and $t_c = 0$ denote the tested degree of association. A single simulation run was carried out by first randomly drawing all θ_{sg} parameters from the representative sampling distributions, until $|T_\theta - t| < 0.001$ was satisfied. This process yields parameters that are both representative and that also exhibit a target degree of association (within a small error margin). Next, the drawn parameters satisfying this condition were plugged into the likelihood of the ZIBR data model, which was in turn used to draw a random sample of relative abundance values. This simulated pseudo-data was then used to fit the Bayesian ZIBR model and conduct our hypothesis test. Each considered scenario was simulated 1000 times and statistical power for given t, n and t_c was then estimated as the proportion of times H_0 (i.e. $T_\theta \leq 0$) was rejected. We specifically considered Pearson correlation as choice for $T(x, y)$ in this simulation.

An additional restriction was put in place for sampling pseudo-data in order to prevent rare cases of sparse datasets with insufficient signal to perform the analysis. If a generated pseudo-dataset contained more than three sub-strata (s, g) in which all observations exhibit a response value of either all 0 or all 1, then it was rejected and a new pseudo-dataset was sampled.

3.4.7 Model fitting

Let j_k denote the unique identifier index for the subject and s_k denote the body site that sample k originated from. Also, let and g_{j_k} denote the disease status for subject j_k, let X_k be the log of total sample abundance for sample k and let b_{j_k} denote the random intercept for subject j_k. The three different models that were utilized in this study are shown below:

$$\text{Model A:} \quad logit(\omega_k) = \beta_{s_k,g_{j_k}} \quad \& \quad logit(p_k) = \beta_{s_k,g_{j_k}}$$

$$\text{Model B:} \quad logit(\omega_k) = \beta_{s_k,g_{j_k}} + b_{j_k} \quad \& \quad logit(p_k) = \beta_{s_k,g_{j_k}} + b_{j_k}$$

$$\text{Model C:} \quad logit(\omega_k) = \beta_{s_k,g_{j_k}} + b_{j_k} \quad \& \quad logit(p_k) = \beta_{1,s_k,g_{j_k}} + X_k\beta_2 + b_{j_k}$$

Model A was utilized in the simulation studies. Model B was utilized for fitting ASV level data, while Model C was utilized for fitting genus level data. This choice was made because scaling OTU counts to relative abundance will only make non-zero relative abundance comparable between samples, but not the rate of absence. This is due to the fact that, even if the true probability of absence p for a specific OTU is very high, if more microbes are overall observed in sample 1 than in sample 2, then the probability of observing none of the microbes belonging to the target OTU in sample 1 is much lower than in sample 2. For example, if a total of 1,000,000 microbes live in a body site and 100 of them belong to the genus *Prevotella*, then if we randomly extract 1,000 microbes from this body site with our sample, we would expect to only rarely find one of these 100 microbes in our sample. However, if our sample randomly extracts 100,000 microbes from the body site, it would be rare to find none of the 100 microbes in it that belong to the genus *Prevotella*. So since the genus level data was not rarefied, the total sample abundance differed between samples and an adjustment was necessary, whereas the ASV level data was rarefied and

did not require adjustment for total sample abundance.

In order to achieve potentially better convergence behaviour and to simplify and speed up the model fitting, the logistic regression component of the model was fit independently of the Beta regression component, in all cases. The resulting posterior chains of p and ω were then used to calculate the posterior chain of ω. This approach is justified under the assumption that p and ω are independent after adjusting for covariates, but may be inadequate when there are confounders affecting both parameters not accounted for in the model.

3.5 Results

3.5.1 Simulation Studies

Performance of our proposed approach was first evaluated using series of simulation studies. In an attempt to obtain sampling distributions of parameters that would approximate biological distributions, unstratified ZIBR models were fit to each OTU in the pancreatic cancer dataset (see Methods for details of this dataset). Unstratified parameter estimates were then used to obtain smooth sampling distributions of ω, p, ϕ. Finally, these sampling distributions were used to generate many pseudo-datasets satisfying H_1 and performance was evaluated when applying the previously described testing approach to the simulated dataset.

Sampling distributions for parameters were similar for both genus and ASV level. However, for ω, the mean non-zero relative abundance, distributions tended to be slightly further concentrated toward 0.0 on the ASV level as compared to the genus level. Further, distributions of p tended to be slightly more concentrated toward 1.0 on the ASV level as compared to the genus level. In both cases a linear relationship was observed between $\log\omega$ and $\log\phi$ which was ultimately used to sample ϕ conditionally on ω (Figure 3.3).

In summary, the following sampling distributions were obtained:

Genus: $p \sim Beta(1.67, 0.4)$; $\omega \sim Beta(0.63, 53.27)$;

$$\log\phi \,|\, \min_{sg}\{\log\omega_{sg}\}) \sim N(-1.02\min_{sg}\{\log\omega_{sg}\} - 1.41, 0.3^2)$$

ASV: $p \sim Beta(7.35, 0.49)$; $\omega \sim Beta(1.46, 121.12)$;

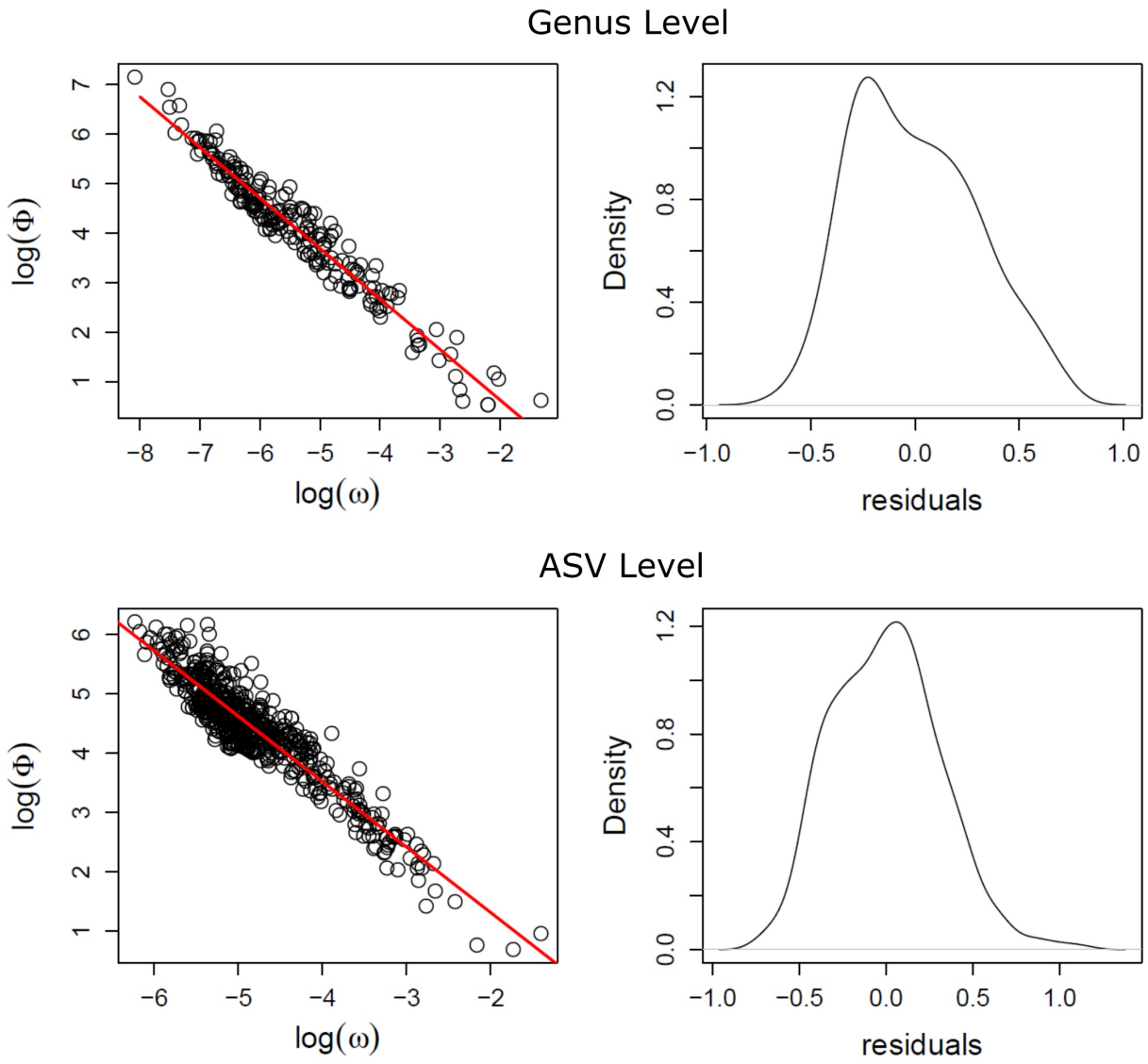


Figure 3.3: Observed relationships between marginal distributions of ω and ϕ estimated from the pancreatic cancer dataset. For both the genus and the ASV level, parameters were estimated marginally for each OTU across all observations without any stratification. When plotting marginal parameter estimates of ω and ϕ a linear relationship can be observed on the log scale. This relationship was utilized to sample ϕ conditionally on ω in the simulation studies.

$$\log\phi\,|\,\min_{sg}\{\log\omega_{sg}\}) \sim N(-1.10\min_{sg}\{\log\omega_{sg}\} - 0.89, 0.31^2)$$

As expected, simulations of biological data revealed that analyses on the genus level were overall more powerful than on the ASV level, regardless of which population parameter was investigated (Figure 3.4). Assuming $t_c = 0$, four disease status groups, 95% credible intervals and utilizing Pearson correlation, the highest power was achieved when testing PASTA of ω. Under a moderate degree of association of $T_\theta = 0.537$ a target power of 0.8 was reached for 5 samples per stratum on the genus level and 15 samples per stratum on the ASV level. Type 1 error rates appeared adequately calibrated to the 5% significance level ranging from 0.03 to 0.056 on the genus level and from 0.032 to 0.06 on the ASV level. Despite the relatively modest within group sample size needed to detect a moderate degree of association for ω with adequate statistical power, there appeared to be considerably less power for tests of p. Under a high degree of association of $T_\theta = 0.834$ a target power of 0.8 was reached for 40 samples per stratum on the genus level. On the ASV level, utilizing as many as 80 samples per stratum resulted in a power of only 0.59 for the same T_θ. Type 1 error rates also appeared mostly calibrated in this scenario, but showed deflation for smaller sample sizes, assuming a value of 0.027 on the genus level and 0.009 on the ASV level.

Testing PASTA of the overall mean $\mu = \omega(1-p)$ was also investigated. While improving with increasing size of effect and sample size, the power for this parameter was lower than when considering ω, p individually. Even when considering the large degree of association $T_\theta = 0.834$ and using 100 samples per stratum, the genus level scenario achieved a power of only 0.546. Notably, type 1 error rates were consistently deflated, ranging from 0.005 to 0.018 on the genus level and 0.002 to 0.008 on the ASV level. Type 1 error rates were deflated across all simulated scenarios, reaching values of less than or equal 0.018 or less.

Discrepancies in performance were found to be directly related to precision of parameter estimates. When plotting the posterior means of T_θ against their true simulated values across various simulation runs, the variation around the identity line consistently increased from ω to p, as well as from genus to ASV level (Figure 3.5). Analogously, posterior distributions of T_θ were found to

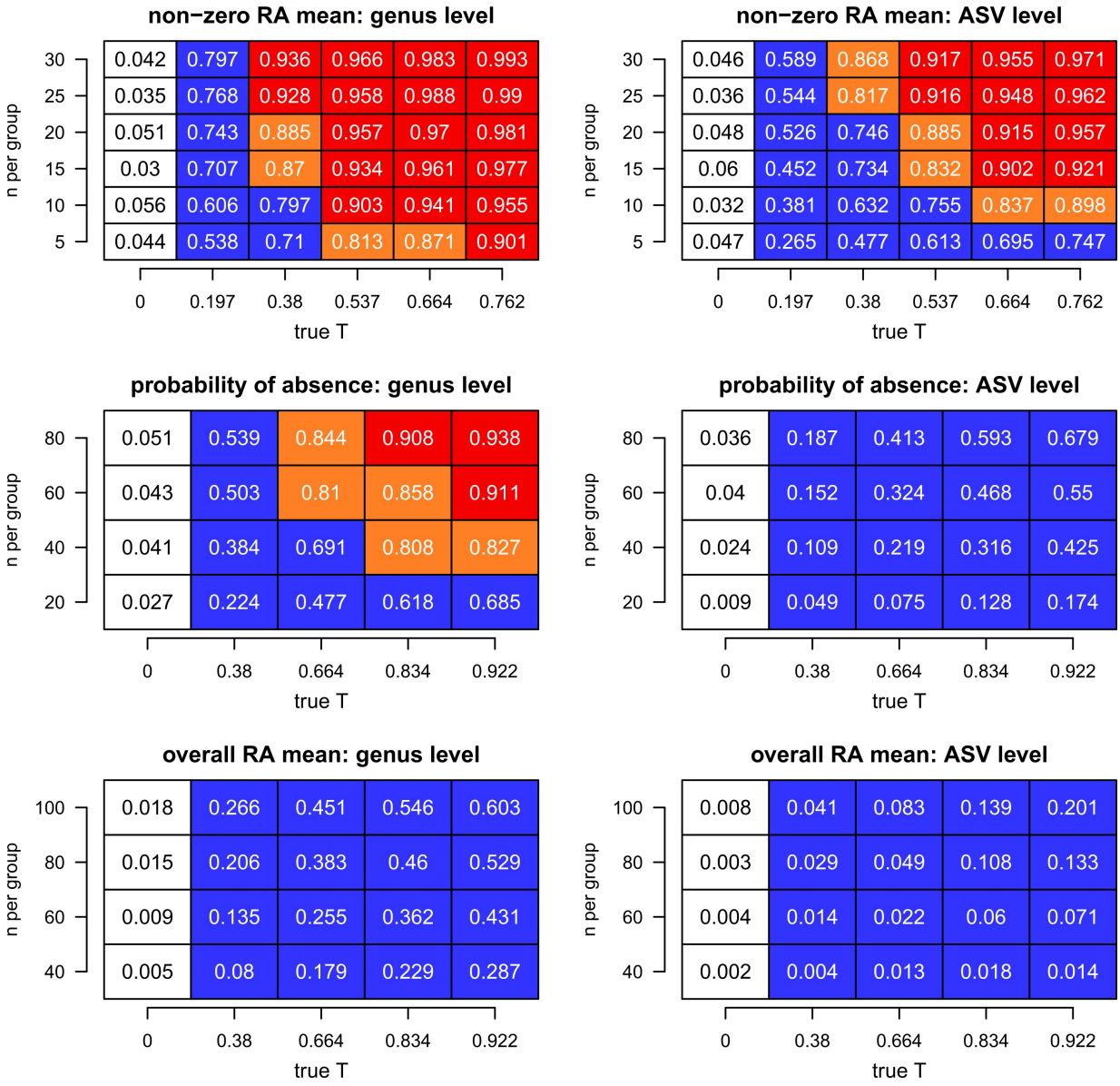


Figure 3.4: Results of the simulation studies. Power plots are displayed for testing PASTA of various population parameters with $t_c = 0$ at both ASV and genus level. The term "n per group" refers to the number of samples available in each of the eight sub-group combinations resulting from two body sites and four different levels of disease status. H_0 was rejected if $Pr(T_\theta|\mathbf{Y} \leq 0) < 0.05$. Type 1 error rates are displayed in white colored boxes with black fonts. Power values less than 0.8 are colored blue, values larger than 0.9 are colored red and values between 0.8 and 0.9 are colored orange. Genus level pseudo data generally has higher statistical power than the ASV level. High performance is achieved by the non-zero mean ω, while an increased sample size is required for the probability of absence p. Tests of the overall mean μ result in low performance, when only mildly constraining sparsity.

on average become more diffuse and more biased towards 0, when moving from ω to p or from genus to ASV level. When performing simulation runs of a scenario with low relative precision, in which p was sampled from a Uniform(0.85,0.95) distribution, the posterior distribution of T_θ was on average almost perfectly centered at zero and highly diffuse.

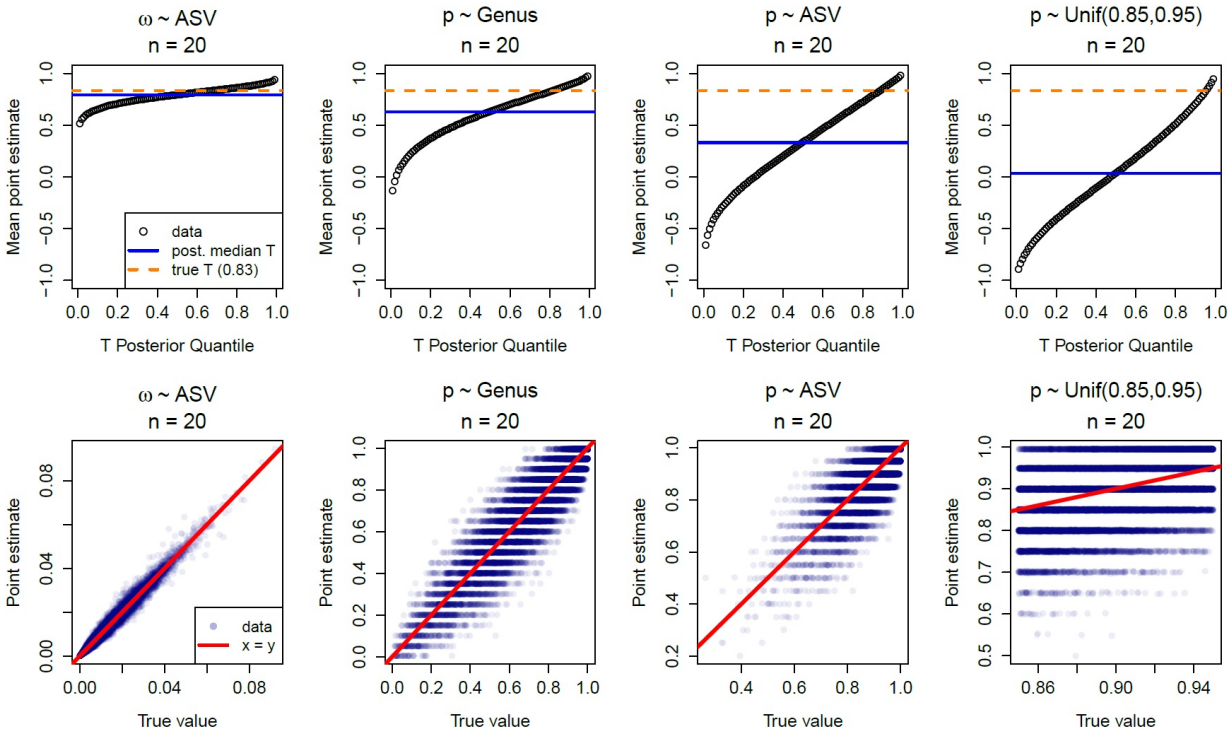


Figure 3.5: Effects of the relative precision of parameter estimates on the posterior distribution of T_θ. The first row shows the average point estimate of posterior quantiles of T_θ across simulation runs for various simulation scenarios. The second row shows the associated plots of the parameters' posterior means versus their true values across simulation runs. As the relative precision of parameter estimates decreases, the posterior distribution of T_θ becomes more diffuse and more biased towards 0.

Poor power when testing μ was also found to be related to two additional factors. Detailed results of simulations accounting for these factors are displayed in Additional File 1. The deflated type 1 error rates when utilizing 95% credible intervals, lead to overly conservative tests that negatively affected power. Calibrating type 1 errors to 5% by adjusting lower bounds of the credible intervals of T_θ for each considered sample size, lead to a consistent improvement in power, reaching a value of 0.73 for $T_\theta = 0.922$ and 80 samples per stratum on the genus level. The second factor that affected performance was the employed liberal three sub-strata rule, which allowed up

to three strata to exhibit exclusively zeroes. Since high rates of absence were simulated, this case often naturally occurred leading to the three respective posterior estimates being imputed with the vague prior distribution, which is very imprecise. A follow-up simulation restricting all strata to have at least one non-zero observation, lead to a consistent increase in power, reaching a value of 0.82 for $T_\theta = 0.922$ and 80 samples per stratum on the genus level. In both settings, overall performance for testing μ was consistently lower than for testing p regardless of taxonomic levels. When both calibrating type 1 error and restricting non-zero observations at the same time, power increased further but was not consistently better than for testing p.

Additionally, three sets of simplified supplementary simulations were also performed to showcase how the PASTA testing approach can be analogously utilized in other data models. A simple Poisson regression model and a log ratio Aitchison model both achieved performance metrics slightly less performant but overall comparable to testing PASTA of the mean non-zero relative abundance via Beta regression. In particular, the Beta regression model appeared to achieve higher power for small sample sizes than the other two approaches and the log-ratio Aitchison model appeared to perform slightly worse than the Poisson regression model. On the other hand, testing PASTA of the overall mean in a zero-inflated Poisson model, utilizing the same smooth sampling distribution of zero-inflation rate p as for ZIBR model on the genus level, achieved performance metrics comparable to testing PASTA of the overall mean in the ZIBR model. Even though minor differences with respect to calibration of type 1 error and statistical power were observed across the different models, the testing approach was overall viable regardless of the scenario. A detailed summary of these simulations is provided in Additional File 2.

3.5.2 Applying the Approach to Biological Data

The 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), is a systematic classification of medical conditions provided by the World Health organization. Disease status information in this dataset was available via ICD-10 codes for each subject. The clinical pancreatic cancer dataset contained four predominant cancer-types:

C24.x, C25.x, K86.2 and other, where ".x" denotes a further sub-type that could differ by subject and "other" refers to pancreatic cancer in various other categories or other diseases of the foregut. OTUs exhibiting significant PASTA with respect to disease status were successfully identified for both genus and ASV level in this dataset. For analysis we considered coding cancer sub-type in two ways: four group coding as described above and three group coding, which collapsed "K86.2" and "other" into one group. On the genus level, when coding disease status into four groups three genera exhibiting PASTA between mouth and gut were identified: *Fusobacterium*, *Haemophilus* and *Veillonella* (Table 3.2). After substratifying oral sites into saliva, tongue, buccal and gum, these association were found to be preserved for some of the site pairs: *Fusobacterium* also exhibited PASTA between gut an saliva sites; *Haemophilus* also exhibited PASTA between gut and gum, as well as gut and tongue; *Veillonella* also exhibited PASTA between gut and gum. Several genera also exhibited PASTA between individual mouth sites (Table 3.3). Two genera exhibited PASTA between four pairs of mouth sites: *Fusobacterium* and *Actinomyces*. Three genera exhibited PASTA between two pairs of mouth sites: *Atopobium*, *Haemophilus* and *Prevotella*. Six genera exhibited PASTA in only one pair of mouth sites.

Table 3.2: Genus level OTUs showing evidence of PASTA between gut and mouth sites when dividing ICD10 code into four groups. For a given genus, a parameter is included in this table if it was marginally significant, or when significance is achieved when T is either Pearson or Spearman correlation. For a given population parameter θ, marginal significance ($Pr(T|\mathbf{Y} \leq 0) < 0.1$) is denoted by $\theta^{\cdot\cdot}$ and significance ($Pr(T|\mathbf{Y} \leq 0) < 0.05$) is denoted by θ^{*}. Three parameters were investigated: μ, ω, p. Due to low power in this exploratory setting multiple testing was not adjusted for.

genus	gut & mouth (all)	gut & buccal	gut & gum	gut & saliva	gut & tongue
Fusobacterium	$\mu^{\cdot\cdot}, p^{\cdot\cdot}$	-	-	$\mu^{\cdot\cdot}$	-
Haemophilus	p^{*}	-	μ^{*}, p^{*}	-	$p^{\cdot\cdot}$
TM7-G1	-	-	-	-	$p^{\cdot\cdot}$
Veillonella	$p^{\cdot\cdot}$	-	$p^{\cdot\cdot}$	-	-

On the ASV level, two ASVs exhibited PASTA with respect to p between mouth and gut when disease status was coded into four groups. When coding disease status into three groups, the same two ASVs as before and three additional ASVs exhibited PASTA with respect to p between mouth

and gut. Notably, among these additional ASVs was a candidate belonging to the *Fusobacterium* genus. Further details of the ASV level analysis are discussed in Chung et al. (2019).

Table 3.3: Genus level OTUs showing evidence of PASTA between mouth sites when dividing ICD10 code into four groups. For a given genus, a parameter is included in this table if it was marginally significant, or when significance is achieved when T is either Pearson or Spearman correlation. For a given population parameter θ, marginal significance ($Pr(T|\mathbf{Y} \leq 0) < 0.1$) is denoted by $\theta^{\cdot\cdot}$ and significance ($Pr(T|\mathbf{Y} \leq 0) < 0.05$) is denoted by θ^{*}. Three parameters were investigated: μ, ω, p. Due to low power in this exploratory setting multiple testing was not adjusted for. Six OTUs showing association for only one pair of mouth sites are not shown in this table.

genus	buccal & gum	buccal & saliva	buccal & tongue	gum & saliva	gum & tongue	saliva & tongue
Actinomyces	-	$\omega^{\cdot\cdot}$	ω^{*}	$\mu^{\cdot\cdot}, p^{*}$	-	μ^{*}
Atopobium	-	-	-	-	p^{*}	$p^{\cdot\cdot}$
Fusobacterium	p^{*}	$\omega^{\cdot\cdot}, p^{\cdot\cdot}$	$\mu^{\cdot\cdot}$	p^{*}	-	-
Haemophilus	-	$\omega^{\cdot\cdot}$	-	-	-	$\mu^{\cdot\cdot}, \omega^{\cdot\cdot}$
Prevotella	-	$\mu^{\cdot\cdot}, \omega^{\cdot\cdot}$	-	-	-	$\mu^{\cdot\cdot}$

3.6 Discussion

The methodology presented in this publication successfully establishes a general framework to test for pairwise stratified association (PASTA) in microbial abundance or relative abundance. The approach first estimates posterior distributions of population parameters $\theta|\mathbf{Y}$ within the strata of body site and disease status and subsequently calculates a correlation statistic T_θ between body sites, which scores their degree of association. This allows researchers to identify individual microbes or groups of microbial species that show consistent abundance patterns between different body sites with respect to the disease status of patients or any other relevant categorical grouping variable.

While this work focuses on identifying preserved patterns between body sites, anti-correlated relationships, where an increase in one body site corresponds to a decrease in another body site, may also be also of biological interest. Such associations are represented by a decreasing functional relationship between the two body sites. Our approach can also be used to identify these relationships by flipping the inequalities in H_0 and H_1 and rejecting the null when $Pr(T_\theta|\mathbf{Y} \geq t_c) < \alpha$. If

either correlated or anti-correlated relationships are to be identified a two-sided test can be analogously formularized testing $H_0 : -t_c \leq T_\theta \leq t_c$.

It has to be noted that while many possible T may be adequate to detect a wide variety of increasing relationships h, the choice of T can favour certain shapes of h. If the Pearson correlation is employed, linear relationships will achieve higher scores than rapid, exponential growth relationships, since it measures the degree of linear association. In this case, overly large values of t_c (for example $t_c = 0.8$) should be avoided as they may lead to falsely rejecting non-linear, increasing relationships. While rank-correlation measures such as the Spearman correlation may be more generally applicable, they may also be less powerful, especially when few groups are considered ($g < 5$). In cases with a small number of groups, the discrete nature of the rank-correlation statistic is more pronounced. When utilizing Spearman correlation it is helpful to keep in mind that T_θ can only assume 4 discrete values when $g = 3$, 11 discrete values when $g = 4$ and 21 possible values when $g = 5$.

Care should also be exercised when interpreting significant associations. The test for PASTA is concerned with trend, agreement or association between $s = 1$ and $s = 2$ after stratification according to g, but does not at all provide information on whether the effect of site s or disease status g is biologically or clinically significant. To the contrary, it assumes that both grouping variables are inherently meaningful objects of the research hypothesis. For example, if there is no significant effect of body site (i.e. abundance is the same between mouth and gut), but abundance differs by disease status, the test statistic will likely score a high degree of association, because what is going on in one site is still associated with what is going on in the other site and this is an inherently meaningful relationship to us. However, the contrary where effect of body site is significant (i.e. abundance is different between mouth and gut) but effect of disease status is not, will not necessarily lead to a significant score of association. Scenarios are possible in which there are small effects of body site and disease status, where none are strong enough to reach statistical significance, yet the test for association may still be overall significant, as long as the trend across strata is pronounced enough. To understand the specific nature of an identified PASTA relationship,

it can be useful to plot credible intervals of parameter estimates θ_{sg} side-by-side (Additional File 3) or to perform statistical follow-up tests investigating the effects of s and g. In order to reduce the burden of multiple testing and to increase the likelihood of screening for impactful associations, a researcher may also choose to first perform marginal tests confirming whether each microbe exhibits significant (or marginally significant) differences with the phenotype of interest within the gut. The restricted set of microbes exhibiting such significant differences could then be used to test for PASTA.

It should be noted that in cases where more information about h is known in advance, more powerful tests of association could be designed that leverage this information. If, for example, h was known to be linear, then the following model could be fit: $\theta_{2g} = \alpha + \theta_{1g} \cdot \beta$, which drastically reduces the number of parameters. In this setting, a PASTA test would be reduced to significance of the parameters α, β. Whilst being more powerful, such a model would also allow one to learn the relationship between mouth and gut, which could be leveraged for predicting gut samples via mouth samples of newly observed subjects. Knowledge about the correlation structure among strata and between OTUs could also potentially be incorporated by utilizing Bayesian hierarchical prediction models with shared hyperpriors. Such sophisticated models may further provide the opportunity to increase power and more adequately reflect knowledge about the data. However, the benefit of our current approach is its general applicability and lack of assumptions about h or correlation structure in the data. Little is currently known about the form of relationships between microbes in different organs or tissues. It is therefore more important to be able to identify cases in which a relationship is present as opposed to fully characterizing the relationship. Without prior knowledge the choice of h is arbitrary and researchers run the risk of potentially missing associations that do not conform with this choice. A researcher can first use our approach to identify microbes exhibiting promising associations, then look at point estimates and credible intervals of parameters across strata to learn about the shape of h. This may then motivate building a prediction that is grounded in empirical evidence. Another benefit of an adequately chosen multivariate Bayesian hierarchical model is that it allows one to test whether OTU A in mouth associates with

OTU B in gut. While such a model has the potential to provide a more powerful test, the here proposed approach does allow one to identify this type of association by including the response values from OTU A in oral samples and the response values of OTU B in gut samples into the model and conducting the test analogously. However, if such a strategy was employed, ϕ may have to be estimated individually per body site, as the assumption of constant dispersion is likely to not hold between different OTUs.

Results of the simulation studies reveal that the testing procedure is able to successfully identify PASTA patterns. The decreased performance on the ASV level can be attributed to decreased signal intensities and the overall increase in sparsity of non-zero observations. The substantial drop in performance when investigating PASTA of p was demonstrated to be a result of overall lower precision in estimation, compared to ω. Since probabilities of absence are generally high and concentrated towards 1.0 across OTUs and strata, the differences between them are often small. In this scenario, to be able to reliably quantify differences and assess trends with adequate precision, larger sample sizes are required. This problem is thus a limitation of the zero-inflated data and not the testing approach itself.

Investigating the overall mean μ, may not always be viable when utilizing the ZIBR model. Since its estimation is based on estimates of both p and ω, its estimates are subject to more sources of variation, resulting in poorer precision and lower power. Our simulation suggests that if the properties of the population that is to be analyzed are well known, adjusting the quantile $t_{Q\alpha}$ to calibrate type 1 error rates is a viable strategy to improve performance. If this was not the case and a researcher was convinced that inference based on μ was more biologically meaningful than considering the individual components p and ω, alternative models may be considered. For relative abundance data an adequate choice may be the marginalized ZIBR model as proposed by Chai et al. (2018) which directly estimates μ as a function of covariates. These estimates could then be used analogously to test for PASTA relationships using the here proposed approach.

The supplementary simulations provided in Additional File 2 should also be interpreted with caution. While their results do provide information about general viability of our testing approach

in the respective scenario, they may not be suitable to infer superiority of either modeling approach. Direct comparison of the models based on the simulation scenarios could be biased, since in each scenario pseudo-data was generated differently and the number of estimated parameters also differed between models.

The fact that in the pancreatic cancer patient dataset OTUs can be identified that show associations between mouth and gut, as well as between individual oral sites suggests that they may be promising candidates for potential biomarkers. Among these were *Fusobacterium* and *Haemophilus*, both oral bacteria recently found to distinguish pancreatic head carcinoma patients from healthy subjects (Lu et al., 2019). Also, species belonging to the genera *Fusobacterium* and *Prevotella* (even though the latter was only found to show association between mouth and gut) have been shown to associate with periodontal disease (Chiranjeevi et al., 2014; Chen et al., 2018). These results lend further credence to the disease related connection between microbial abundance in mouth and gut and suggests that our method leads to conclusions consistent with the literature. More future research will be needed to validate these findings.

The simulation studies also confirmed that tests of PASTA applied to the pancreatic cancer patient dataset are likely underpowered due to the limited sample size. It should be noted that many OTUs could not be tested due to too high zero-inflation and thus insufficient signal. These two factors likely explain why relatively few candidates were identified when conducting the tests. Future studies may consider larger sample sizes or aim to improve the yield of observed counts in each sample to alleviate this issue. Our results suggest that differences in the extent of zero-inflation between groups may be generally hard to detect for small to medium sized studies when more granular phylogenetic levels are targeted.

3.7 Conclusions

In conclusion, the performed simulation studies demonstrate the viability of the approach in the context of ZIBR models and suggest that for tests of association of mean non-zero relative abundance modest sample sizes can achieve adequate power for moderate degree of association. The

simulations also highlight potential lack of power for low-level phylogeny data (e.g., species, ASV) or when more complex functions of population parameters are considered. When analyzing a biological dataset consisting of pancreatic cancer patients the approach is able to identify microbes that exhibit PASTA patterns and are consistent with independent findings of current research studies. The generality of this approach allows it to be extended to other data models and research settings, ensuring that it can be useful for researchers interested in stratified associations in the microbiome world and beyond.

Chapter 4

Leveraging Spatial Correlation to Improve Analysis of Cell Type Specific Methylation from Whole Blood

4.1 Statement of Contributions

In this project, I, Richard Meier, developed the methodology, conducted the simulation studies and statistical analyses, and wrote the manuscript.

4.2 Abstract

Methylation of Cytosine-Guanine dinucleotides (CpGs) in mammals plays an important role in the regulation of cellular processes and change in methylation has been linked to many human diseases. Cell type specific analysis of methylation levels in biological samples has shown promise to improve insight into the interplay of cellular processes and disease. Recent statistical modelling strategies of Zheng et al. (2018) and Rahmani et al. (2019) now allow for these analyses to be performed based on bulk methylation data, without the necessity of isolating cells. Unfortunately, none of these approaches incorporate the knowledge that CpGs tend to spatially correlate with base-pair distance. Here, we present a Bayesian hierarchical modelling strategy that leverages spatial correlation in order to improve model fit and statistical power when testing for cell type specific differential methylation. Whole blood methylation data of isolated cell types is first empirically evaluated to motivate the approach and subsequently utilized in extensive simulation studies comparing benefits of candidate models Our approach consistently improved prediction accuracy and statistical power compared to non-spatial models, being particularly beneficial when data was noisy or lowly abundant cell types were present. Our results suggest that, future studies of cell-type specific methylation based on bulk samples that utilize our approach will be more efficient and require smaller sample sizes than previous approaches.

4.3 Introduction

DNA methylation (DNAm) patterns play a key role in how other molecules interact with the DNA molecule and are involved in fundamental biological processes such as gene silencing, tissue differentiation and cellular development (Moarii et al., 2015). Changes in DNAm affect cellular phenotypes (Zheng et al., 2018) and have been linked to many types of cancer, autoimmune diseases, metabolic disorders, neurological disorders and aging (Jin & Liu, 2018). In humans, one of the most important types of methylation is targeting so called CpGs (i.e. CpG dinucleotides), referring to genomic locations in which a cytosine base is immediately followed by a guanine base, in particular when moving in the direction from the 5' end to the 3' end of the DNA molecule. Methylation of CpGs refers to the attachment of a methyl group to the cytosine via specific enzymes. A major challenge in the analysis of DNAm data is the fact that methylation is often profiled in heterogeneous biospecimens comprised of ensembles of different cell types (e.g. peripheral whole-blood, tumor tissue, etc.), each assuming unique functions and exhibiting distinct, epigenetic patterns (Christensen et al., 2009; Koestler et al., 2012; Reinius et al., 2012). This implies that cell type specific analyses of differential methylation have a greater potential for identifying cellular mechanisms of disease (Rahmani et al., 2019). Two key challenges in the analysis of DNAm in bulk tissue are the potential for confounding by cell heterogeneity (Teschendorff & Zheng, 2017) and that bulk analyses of heterogeneous samples may conceal true differences restricted to a specific cell type (Zheng et al., 2018; Rahmani et al., 2019). With respect to the latter, this is especially true if cell types constitute a small proportion of the cellular landscape underlying the profiled biospecimen.

Cell specific methylation signatures can be obtained by either cell sorting techniques followed by methylation profiling in specific cell populations or single cell sequencing; both approaches are currently "drastically restricted in their sample sizes owing to high costs and technical limitations" (Rahmani et al., 2019, p. 2). As Rahmani et al. (2019) aptly point out, regardless of future advances in these areas, the current availability of both large-scale bulk methylation datasets and large quantities of bulk samples in general, further motivate the need for statistical methods able to

identify cell-specific effects of methylation based on bulk data.

Recently, two novel statistical methodologies have emerged that are able to perform cell type specific analyses based on heterogeneous bulk samples (Zheng et al., 2018; Rahmani et al., 2019). Both approaches effectively utilize linear models with interaction terms between effects of covariates and proportions of cell type abundance in order to estimate cell-type specific methylation. This type of approach is especially promising, since cell type proportions can be estimated from bulk data with high precision via publicly available reference methylation signatures (Salas et al., 2018; Koestler et al., 2016; Houseman et al., 2012). Unfortunately, neither approach accounts for the spatial correlation present in the methylation signature of proximal genomic CpG loci (Liu et al., 2014; Affinito et al., 2020), which may lead to overestimation of uncertainty and underpowered statistical tests. Furthermore, leveraging spatial correlation has already been successfully utilized in methods that identify differentially methylated regions (containing multiple CpGs) that associate with phenotypes of interest in bulk samples (Jaffe et al., 2012; Catoni et al., 2018). Thus, incorporating this spatial correlation structure into the data model could potentially improve estimation and may lead to more powerful statistical tests of cell-specific differential methylation.

Building on the likelihood modelling structure of Rahmani et al. (2019), we propose a Bayesian hierarchical regression model that leverages the spatial correlation in methylation of nearby CpGs by shrinking estimates of mean methylation towards each other within clusters of nearby CpG loci. The sample correlation structure of a biological dataset containing DNAm signatures profiled in whole blood samples collected from healthy donors is explored, and based on these results an algorithm for grouping CpGs into clusters is formulated. Utilizing this algorithm, a variety of different hierarchical candidate models are proposed and fit to each cluster. Extensive simulation studies are utilized to compare candidate models and to explore benefits of the approach in terms of model fit, prediction, and statistical power. In this work, we focus on analyzing blood derived DNAm data to detect cell-specific and CpG specific differential methylation with respect to some phenotype or exposure of interest.

4.4 Methods

4.4.1 CpG Methylation and Cell Type Deconvolution

Methylation levels can be assessed through assaying techniques that measure how often methylated and unmethylated copies of a target CpG are observed in a given sample. For the Illumina HumanMethylation BeadArrays, this information is commonly summarized for each CpG in each sample as the so called "beta value" which is the ratio between methylated signal intensities, and the sum of methylated and unmethylated signal intensities. This means that the data of interest collected in epigenome-wide association studies (EWAS) will usually be a matrix containing one row for each sample and one column for each CpG. The total CpGs may vary according to the assay used for assessment of DNAm, but is generally large, reaching a value of 850,000 in the Illumina Infinium MethylationEPIC array, the most commonly used technology for the assessment of methylation in large-scale epidemiological studies in humans.

The fundamental involvement od DNAm in processes on the cellular level results in the emergence of stable, cell type specific methylation patterns (Christensen et al., 2009; Koestler et al., 2012; Reinius et al., 2012). This means that by isolating cell types from a specific tissue or biospecimen (e.g. whole blood), it is possible to construct a reference matrix $\mathbf{R}$ of stable beta values for a specific subset of CpGs (rows of the matrix) and for each cell type (columns of the matrix), such that patterns in the matrix can distinguish cell types with high precision (Houseman et al., 2012). Assume such a reference matrix contains all major cell types occurring in a specific biospecimen and is also representative of said biospecimen. Further, let the vector $\mathbf{W_i}$ denote the cell proportions of a given sample i. Every element W_{hi} in this vector contains the proportion of the biological landscape found in sample i that belongs to cell type h. $\mathbf{W_i}$ can then be estimated from the observed bulk methylation beta values $\mathbf{X_i}$ of CpGs via the following linear relationship: $E[\mathbf{X_i}] = \mathbf{RW_i}$ (Houseman et al., 2012; Titus et al., 2017; Salas et al., 2018). Such deconvolution methods allow for the estimation of cell proportions, required for cell type specific analyses, even if no explicit cell counting technique was employed.

4.4.2 Biological Dataset

Exploratory evaluation of spatial correlation and simulation studies, both make use of a biological dataset first published by Salas et al. (2018), containing methylation beta values from samples of six blood cell types isolated from anonymous, healthy donors. The following blood cell types were isolated via fluorescence activated cell sorting (FACS): neutrophils, monocytes, B-lymphocytes, natural killer cells, CD4+ T-cells and CD8+ T-cells. Methylation levels in this dataset were obtained using the Illumina HumanMethylationEPIC array, as well as the "minfi" (Aryee et al., 2014; Fortin et al., 2016) and "EnMIX" (Xu et al., 2015) R packages for preprocessing and data quality control. Further details about the sample preparation and collection protocols are provided in Salas et al. (2018).

4.4.3 Overview of the Approach

The approach proposed in this study aims to improve statistical model performance by utilizing the knowledge that methylation values tend to spatially correlate among CpGs in close proximity. Introducing this information into the model provides the opportunity for methylation estimates of proximal CpGs to borrow information from each other, potentially resulting in smaller credibility intervals and more accurate estimates. We hypothesize that this spatial correlation does not stem from a genome-wide or chromosome-wide correlation function shared among CpGs which decreases as distance between CpGs increases. Rather, we hypothesize that spatial relationships can be adequately captured and approximated by considering clusters of nearby CpGs. Our model assumes that on any given chromosome there are likely many different clusters of nearby CpGs, varying in total CpG number and spanned distance, such that within each cluster there is a different degree of positive methylation level correlation among CpGs. Further, we assume that there are also likely hierarchies in which clusters can correlate with each other, provided they are not spaced too far apart.

Inspired by the multi-level hierarchical model proposed by Berry & Berry (2004), spatial relationships are thus translated into a hierarchy in which CpGs form clusters and multiple clusters

form so called super clusters. The model then assumes that CpGs within a cluster correlate most closely and CpGs in different clusters but within the same super cluster correlate less strongly. This is achieved by shrinking estimates of mean methylation within the same cluster towards a common, overall cluster mean and then in turn shrinking cluster methylation means towards a common super cluster mean. A conceptual figure of the modeling structure is provided in Figure 4.1.

Our approach was applied as follows: First, autocorrelation of DNAm as a function of base-pair distance was empirically evaluated in the biological whole blood methylation dataset. Based on these results, distance thresholds were devised that were utilized in an algorithm that grouped proximal CpGs into clusters and in turn proximal clusters into super clusters. Bayesian hierarchical regression models were then separately fit to each super cluster.

Evaluation of this modelling strategy was achieved via a variety of extensive simulation studies based on the biological dataset in which performance of candidate models was compared.

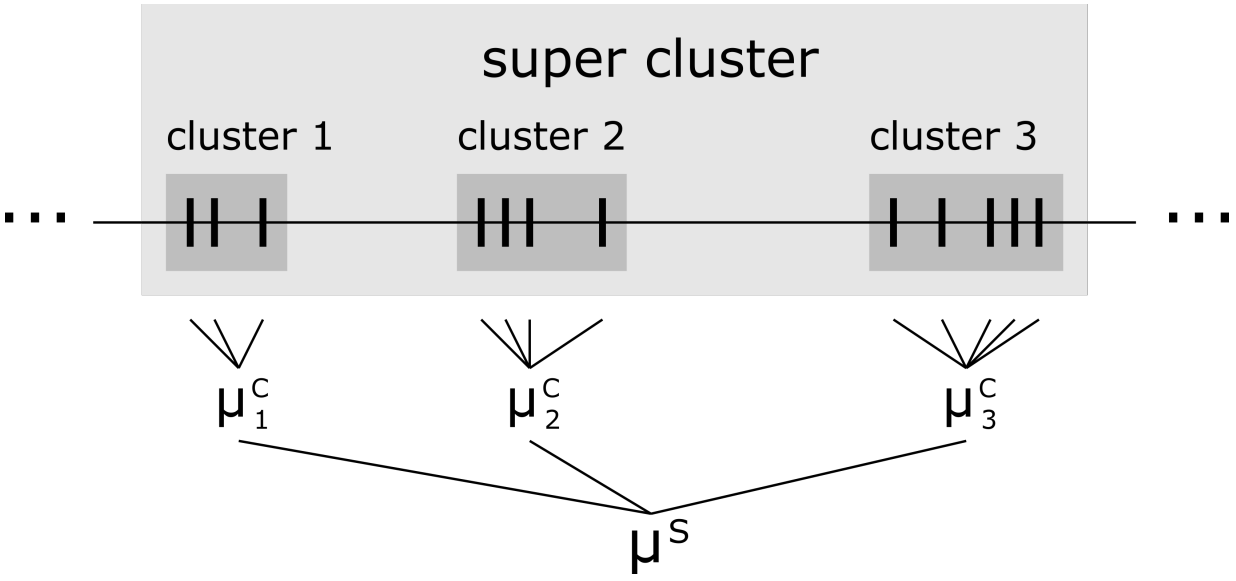


Figure 4.1: Conceptual overview of the spatial model structure. Displayed is a snapshot of a hypothetical chromosome, represented by a horizontal axis line. Mean methylation levels of CpGs are denoted as vertical tic marks along the axis line. Mean methylation values within the same cluster k are shrunk towards the overall cluster mean μ_k^C and cluster means are in turn shrunk towards the overall super cluster mean μ^S.

4.4.4 Emprical Analysis of DNAm Autocorrelation

For each individual cell type, Pearson correlation of sample beta values was evaluated as a function of base-pair distance in a sliding window strategy. This strategy was employed instead of calculating the complete correlation matrix of all CpGs on a chromosome, because the sheer number loci would make this calculation intractable and also because correlation between highly distant loci was expected to be low and not of primary interest. After choosing a target chromosome and a random CpG starting index, correlations were calculated for the 50 neighboring CpGs downstream of the starting index. The off-diagonal upper triangle of the correlation matrix was then stored and the same calculation was performed for the next 50 downstream CpGs. This was repeated until a total of 1800 CpGs were processed. All extracted correlation values were then plotted against base-pair distance as two-dimensional heatmaps and a LOESS smoothed trend was overlaid via the "ggplot2" (Wickham, 2016) R-package.

4.4.5 Definition of Models

All models considered in this study utilize the same basic likelihood function but differ in their structure and assignment of prior distributions. This likelihood was adopted from the Tensor Composition Analysis (TCA) model by Rahmani et al. (2019), which achieved the current best reported performance for cell-type specific analyses based on bulk methylation data. In contrast to the TCA approach, our approach considers this likelihood for only a specific set of clusters of CpGs, i.e. a target super cluster, at a time and does not consider the entire methylation matrix of a particular chromosome at once. Let Z_{ihj} denote the cell type specific methylation beta value of cell type $h = 1, 2 \ldots, H$, where H is the number of cell types, and CpG $j = 1, 2, \ldots, J$, where J is the number of CpGs, in sample i. Furthermore, let $X_{i,j}$ denote the bulk methylation beta value for CpG j in

sample i. The two-part likelihood of the data is then defined as follows:

$$Z_{ihj} = \mathbf{c_i}^T \delta_{hj} + \varepsilon_{ihj} \quad \text{where} \quad \varepsilon_{ihj} \overset{iid}{\sim} N(0, \psi_{hj}^{-1})$$
$$X_{i,j} = \sum_{h=1}^{H} w_{hi} Z_{ihj} + \kappa_{ij} \quad \text{where} \quad \kappa_{ij} \overset{iid}{\sim} N(0, \tau^{-1})$$

In this framework, the vector $\mathbf{c_i}$ contains all cell type specific covariates for sample i (e.g. intercept, age, gender ...etc.) and the parameter vector δ_{hj} captures effects of covariates on cell-type specific methylation. Lastly, w_{hi} denotes the proportion of cells of type h in sample i, satisfying $0 \leq w_{hi} \leq 1$ and $\sum_{h=1}^{H} w_{hi} = 1$.

While the type of model proposed in this paper is in principle applicable to any type of multiple regression scenario with several covariates, we consider two simple experimental designs in particular:

1. The marginal design: $Z_{ihj} = \delta_{0hj} + \varepsilon_{ihj}$, in which all observations are assumed to originate from cell types with the same group means

2. The two arm design: $Z_{ihj} = \delta_{0hj} + \delta_{1hj} \cdot I[i \in TRT] + \varepsilon_{ihj}$, in which observations are assumed to originate from two distinct treatment arms and in which $I[i \in TRT]$ denotes an indicator function that equals 1.0 when observation i is in the treatment arm and equals 0.0 if it is in the control arm instead.

Regardless of the study design, several classes of models were employed to estimate δ. First, we consider a class of models that do not incorporate spatial correlation and are most similar to the TCA approach among the chosen candidate models. Let t denote any given covariate utilized in the model, and let this class of models be denoted as "TM". This class is then defined by the above likelihood while treating all δ as fixed effects and utilizing weakly informative priors for them:

$$\pi(\delta_{thj}) \sim N(0, 100)$$

The first class of introduced spatial candidate models, which we will denote as "SCM1", intro-

duces a separate hierarchical structure for each individual covariate t, when estimating its effects on mean methylation. To define this structure, let $L(j) = k$ represent the cluster index that a CpG j belongs to. On the lowest level of the hierarchy, effect estimates of CpGs $\gamma_{t,h,j}$ are shrunk towards the mean effect $\gamma^{[O]}_{t,h,L(j)}$ of all CpGs within the same cluster. Next, each mean cluster effect $\gamma^{[O]}_{t,h,k}$ is shrunk towards the mean super cluster effect $\gamma^{[OO]}_{t,h}$. Finally, super cluster effects are shrunk towards the overall mean covariate effect on methylation status across all cell types $\gamma^{[OOO]}_{t}$. This multi-step process allows estimates of covariate effects to borrow information from neighboring CpGs according to their degree of spatial correlation. The full model can be expressed as follows:

$$\begin{aligned}
\delta_{t,h,j} &= \gamma_{t,h,j} \cdot (1 - I[L(j)]) + \gamma^{[O]}_{t,h,L(j)} \cdot I[L(j)] \\
\pi(\gamma_{thj}) \sim N\left(\gamma^{[O]}_{t,h,L(j)}, \xi^{[O]}_{t}\right) \quad &; \quad \pi\left(\gamma^{[O]}_{t,h,k}\right) \sim N\left(\gamma^{[OO]}_{t,h}, \xi^{[OO]}_{t}\right) \\
\pi\left(\gamma^{[OO]}_{t,h}\right) \sim N\left(\gamma^{[OOO]}_{t}, \xi^{[OOO]}_{t}\right) \quad &; \quad \pi\left(\gamma^{[OOO]}_{t}\right) \sim N(0, 100) \\
\pi\left(\xi^{[O]}_{t}\right), \pi\left(\xi^{[OO]}_{t}\right), \pi\left(\xi^{[OOO]}_{t}\right) &\sim Gamma(0.1, 0.1)
\end{aligned}$$

Here the indicator function $I[k]$ returns a value of 1 if a cluster k is degenerate, i.e. it only contains a single CpG, and returns a value of 0 otherwise. The purpose of the above reparameterization of $\delta_{t,h,j}$ is to alleviate parameter redundancy in the presence of degenerate clusters. We define a cluster to be degenerate when it only contains one single CpG. Without the reparameterization $\gamma_{t,h,j}$ and $\gamma^{[O]}_{t,h,L(j)}$ would otherwise coincide.

The second class of spatial correlation models, denoted as "SCM2", follows the same motivation as the "SCM1" model but does not assume any relationship of covariate effects between cell types. It is defined as follows:

$$\begin{aligned}
\delta_{t,h,j} &= \gamma_{t,h,j} \cdot (1 - I[L(j)]) + \gamma^{[O]}_{t,h,L(j)} \cdot I[L(j)] \\
\pi(\gamma_{thj}) \sim N\left(\gamma^{[O]}_{t,h,L(j)}, \xi^{[O]}_{t}\right) \quad &; \quad \pi\left(\gamma^{[O]}_{t,h,k}\right) \sim N\left(\gamma^{[OO]}_{t,h}, \xi^{[OO]}_{t}\right) \\
\pi\left(\gamma^{[OO]}_{t,h}\right) \sim N(0, 100) \quad &; \quad \pi\left(\xi^{[O]}_{t}\right), \pi\left(\xi^{[OO]}_{t}\right) \sim Gamma(0.1, 0.1)
\end{aligned}$$

Lastly, the class denoted as "SCM3" dismisses the multi-level hierarchy of "SCM2" and simply directly shrinks each covariate effect towards the overall mean covariate effect of the super cluster. This model can be interpreted as treating each super cluster simply as a large cluster in which all CpGs are equally correlated, without any nuance about closer proximity. It is defined as follows:

$$\begin{aligned}
\delta_{t,h,j} &= \gamma_{t,h}^{[OO]} \\
\pi(\gamma_{thj}) &\sim N\left(\gamma_{t,h}^{[OO]}, \xi_t^{[OO]}\right) \\
\pi\left(\gamma_{t,h}^{[OO]}\right) \sim N(0,100) \quad &; \quad \pi\left(\xi_t^{[OO]}\right) \sim Gamma(0.1, 0.1)
\end{aligned}$$

Each of the four sets of models was considered in three different settings, utilizing three types of prior distributions for the variance parameters of beta values:

- weakly informative: $\pi(\tau), \pi(\psi_{hj}) \sim Gamma(0.01, 0.01)$
- informative type 1: $\pi(\tau), \pi(\psi_{hj}) \sim Gamma(1.36, 0.01)$
- informative type 2: $\pi(\tau), \pi(\psi_{hj}) \sim Gamma(0.844, 0.001)$

The key difference between the weakly informative and the informative priors is that the former is agnostic about the restriction of the parameter space of the variance when response values are bounded between 0 and 1, whereas the latter favors posterior densities of precision parameters that concentrate at values larger than 4. This is desirable, since the variance of any random variable bounded by the (0,1) interval can never exceed 0.25, which is equivalent to stating that the precision parameter for such a variable cannot be smaller than 4. More specifically, parameters of both informative Gamma priors were chosen such that the probability to draw a precision value smaller than 4 would be approximately 1%. The key difference between the two distributions is the anticipated range into which values of variances will most likely fall. Type 1 informative priors exhibit a 99% probability to draw precision values smaller than 538.7, which translates to favoring variances that are larger than 0.002. On the other hand, type 2 informative priors are more open to

considering smaller variances, exhibiting a 99% probability to draw precision values smaller than 4237.8, which translates to favoring variances that are larger than 0.0002.

4.4.6 Estimation and Model Fit Characteristics

Posterior distributions of parameters were estimated using the OpenBUGS (version 3.2.3 rev 1012) software interfacing with the R (version 3.4.0) package "R2OpenBUGS" (version 3.2.3.2), which employed a Markov-Chain Monte Carlo sampling technique to draw 15000 posterior samples of parameters after discarding 1000 burn-in iterations. Model fit was evaluated based on two primary characteristics: model complexity and mean posterior prediction error.

Model complexity was evaluated via the estimated effective number of parameters p_D, as provided by OpenBUGs. Briefly, for a hypothetical response vector $\mathbf{Y}$ with likelihood $f(\mathbf{y}|\theta)$ this commonly employed diagnostic is defined as: $p_D = 2\log f(\mathbf{y}|E[\theta|\mathbf{Y}]) - 2E[\log f(\mathbf{Y}|(\theta|\mathbf{Y}))]$.

Prediction error was evaluated by utilizing two separate datasets: a training dataset, which was used to estimate posterior distributions of parameters, and a testing dataset that was withheld during estimation. Let $\mathbf{X}$ denote the response values of the training data, let $\mathbf{Q}$ denote the response values of the testing data and let $\hat{\mathbf{Q}}(\theta|\mathbf{X})$ denote the random vector following the posterior predictive distribution of $\mathbf{Q}$ given the posterior distribution of the parameters θ based on the training data. The root mean square prediction error (RMSPE) of $\mathbf{Q}$ can then be defined as: $RMSPE(\mathbf{Q}|\mathbf{X}) = \sqrt{\frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}(\hat{Q}_{ij}(\theta|\mathbf{X}) - Q_{ij})^2}$. Since this quantity is itself a random variable, it is summarized by its mean $\widehat{RMSPE}_Q = E[RMSPE(\mathbf{Q}|\mathbf{X})]$, which we will refer to as posterior mean RMSPE. Generally, this quantity will be larger for candidates with worse overall model fit and smaller for candidates with better overall model fit. Since its calculation is based on independent testing data, this characteristic is robust to overfitting.

4.4.7 Simulation Studies

4.4.7.1 Outline

Three major types of simulation studies were performed, each with a different aim. The first set of simulations, which we will denote as type A, aims to identify the most suitable candidate in terms of model fit characteristics across the previously described set of candidate models. After the target model is chosen based on pD and $\widehat{RMSPE}_Q$, simulations of type B aim to further explore model performance in terms of conducting hypothesis tests that try to identify mean differences in methylation of individual CpGs between two hypothetical treatment arms. Lastly, simulations of type C explore benefits of the proposed modeling approach for cell types occurring with different proportions of abundance. This is of interest since if a cell type exhibits consistently smaller cell proportions than other cell types in the same biospecimen, then effects of covariates on DNAm for this cell type are expected to be estimated with lower precision and tests of these effects are expected to exhibit lower statistical power. Due to time constraints and computing limitations, simulations of type A and B were manually terminated after running for two weeks without interruption while simulation C was manually terminated after running for one week without interruption.

4.4.7.2 Data Generating Process for Simulation A

First, the biological dataset is subset to all CpGs originating from a target chromosome and all CpGs are grouped into clusters, which in turn are grouped into super clusters. The cluster generating algorithm employed to achieve this task is further described in section 4.4.8.

Let $M = M(s)$ be a shorthand, denoting the number of CpGs a super cluster s contains and let $\alpha = (2,2,2,2,2,2)$ denote the parameter vector of a utilized Dirichlet distribution. Then for each super cluster s the following steps are performed:

1. Consider a hypothetical treatment and control arm between which mean beta values are to be compared for each CpG and cell type. Let $\mathbf{f}^{[\mathbf{T}]}$ denote a vector of signs of the group mean difference between the two arms, which contains a separate value for each of the six cell

types and is assigned to to be a random permutation of the set $(-1,-1,-1,1,1,1)$.

2. For each subject, draw six cell fractions from a *Dirichlet* (α) distribution yielding the $\mathbf{W}^{[\mathbf{Z}]}$ matrix, containing six columns for each cell type and one row for each subject in the training dataset.

3. For each subject, draw six additional cell fractions from a *Dirichlet* (α) distribution analogously, yielding the $\mathbf{W}^{[\mathbf{Q}]}$ matrix, containing six columns for each cell type and one row for each subject in the testing dataset.

4. For each cell type h, draw a random sample from the pool of biological, cell sorted samples that originate from cell type h and then extract its beta values for each CpG q in super cluster s, yielding the vector of cell type specific intercept terms $(\mu_h)_{Mx1}$.

5. For each subject i, draw cell type specific beta values Z_{ihq} from a normal distribution with the following parameters: $Z_{ihq} \sim N\left(\mu_{hq}, (\sigma^{[B]})^2\right)$

6. For each subject i, draw bulk beta values from a normal distribution with the following parameters: $X_{iq} \sim N\left(\mathbf{Z_i}\mathbf{W_i^{[Z]}}, (\sigma^{[B]})^2\right)$, where $\mathbf{Z_i}$ denotes the matrix of cell type specific beta values with six rows for each cell type and M columns for each CpG and $\mathbf{W_i^{[Z]}}$ denotes the vector of six cell fractions for subject i.

7. Analogously to step 5. and 6., draw another independent bulk methylation dataset $\mathbf{Q}$ using $\mathbf{W}^{[\mathbf{Q}]}$, for the purpose of calculating the posterior mean RMSPE.

4.4.7.3 Data Generating Process for Simulation B & C

Let $L(j)$ denote the cluster index as defined in section 4.4.5, let $M = M(s)$ be defined as previously and let Δ denote the overall effect size of differential methylation between two treatment arms. The key difference between simulation B and C is the choice of the parameter vector α. In simulation B, $\alpha = (2,2,2,2,2,2)$ is assigned the same way as in simulation A, representing cell proportions varying at random. In simulation C, $\alpha = (6.4, 3.2, 1.6, 0.8, 0.4, 0.2)$ is assigned in order to create

cell proportions that consistently decrease from cell type to cell type, but that also maintain a similar degree of overall variability. The latter is ensured by both parameter vectors achieving similar total sums, i.e. 12 and 12.6 respectively. For these two types of simulations (B and C), the data generating process is performed identically to type A up to step 3. and then proceeds in the following way:

4. Let $\Omega^{(s)}$ denote the correlation matrix between effect sizes within cell type and across the M CpGs. Considering any given pair of two CpGs q and r in the same super cluster, this matrix is then constructed as follows...

$$\Omega_{qr}^{(s)} = \begin{cases} 1 & \text{if } q = r \\ \rho^{[C]} & \text{if } q \neq r \cap L(q) = L(r) \\ \rho^{[C]}/2 & \text{otherwise} \end{cases}$$

5. For each cell type h, draw M effect sizes from a $MVN\big(\text{rep}(\Delta, M)\,,\,(\sigma^{[C]})^2 \cdot \Omega_{MxM}\big)$ distribution, yielding the cell type specific effect size vector $(\delta_h^{(s)})_{Mx1}$.

6. For each cell type h, draw a random sample from the pool of biological, cell sorted samples that originate from cell type h and then extract its beta values for each CpG q in super cluster s, yielding the vector of cell type specific, empirical intercept terms $(\mu_h)_{Mx1}$.

7. For each subject i, draw cell type specific beta values Z_{ihq} from a normal distribution with the following parameters: $Z_{ihq} \sim N\left(\mu_{hq} + I[i \in TRT] \cdot \delta_{hq}^{[s]} \cdot f_h^{[T]}\,,\,(\sigma^{[B]})^2\right)$, where $I[i \in TRT]$ denotes an indicator function that equals 1.0 when subject i is in the treatment arm and equals 0.0 otherwise.

8. For each subject i, draw bulk beta values from a normal distribution with the following parameters: $X_{iq} \sim N\left(\mathbf{Z_i}\mathbf{W_i^{[Z]}}\,,\,(\sigma^{[B]})^2\right)$, where $\mathbf{Z_i}$ denotes the matrix of cell type specific beta values with six rows for each cell type and M columns for each CpG and $\mathbf{W_i^{[Z]}}$ denotes the vector of six cell fractions for subject i.

9. Analogously to step 7. and 8., draw another independent bulk methylation dataset **Q** using $\mathbf{W}^{[\mathbf{Q}]}$, for the purpose of evaluating the prediction error on a new, previously unobserved dataset.

4.4.7.4 Power Analysis

For a given cell type h and CpG j the following were the hypotheses of interest:

$H_0 : \delta_{hj} = 0$, i.e. there is NO difference in mean methylation between the two treatment arms

$H_1 : \delta_{hj} \neq 0$, i.e. there IS a difference in mean methylation between the two treatment arms

Simulation B and C primarily explored the ability of models to detect differentially methylated CpGs within cell types in the two-arm design setting. Since each model jointly estimated parameters for multiple cell types and multiple CpGs, testing differential methylation was considered on two levels: The overall performance level across all cell types and across all CpGs and the cell type specific performance level, which considered performance individually within each cell type but across all CpGs.

The overall power and type 1 error were estimated as follows: In each iteration, for a given model and a given rejection rule calculate the proportion of CpGs for which H_0 was rejected across all cell types and all CpG loci in the current super cluster. This proportion can be understood as a preliminary rejection rate estimate based on a single model fit. If, for example, a model considers a total of $H \cdot J = 6 \cdot 4 = 24$ cell type specific CpGs out of which 11 are rejected, then the preliminary estimate for this iteration will be $11/24 = 0.458$. After these preliminary rejection rates are calculated for every simulation iteration, they are averaged in order to yield the final estimate of the overall rejection rate. This rate represents the type 1 error rate when data is simulated under H_0 and the statistical power when data is simulated under H_1.

The cell type specific power and type 1 error were estimated as follows: In each iteration, for a given model, a given rejection rule and a given cell type calculate the proportion of CpGs for which H_0 is rejected. This will yield one preliminary rejection rate estimate for each of the 6 individual

cell types for each simulation iteration. For each individual cell type, average the corresponding preliminary rejection rates across all iterations, yielding the final estimate of cell type specific rejection rate.

4.4.8 Cluster Generating Algorithm

In the following paragraphs a step-by-step description of the cluster generating algorithm is provided. The algorithm utilizes several parameters ($\zeta_1, \zeta_2, \Upsilon_1, \Upsilon_2$) that control the size and span of clusters and super clusters. Υ_1 directly controls the maximum number CpGs that are allowed within any given cluster, while Υ_2 directly controls the maximum number of clusters that are allowed within any given super cluster. These artificial size limits are imposed in order to prevent the formation of large clusters or super clusters that contain a large total number of CpGs. Such large clusters are primarily problematic, because the required run time and computer memory of Bayesian MCMC model fits that are applied to super clusters will increase rapidly as the total number of CpGs increases. By assigning $\Upsilon_1 = 10$ and $\Upsilon_2 = 8$, the total number of CpGs per super cluster is limited to $\Upsilon_1 \cdot \Upsilon_2 = 80$, which could still be processed by the computers employed in the simulation study. Υ_2 was assigned to be slightly lower than Υ_1 to put an emphasis on effects of local structures, since closely proximal CpGs are expected to be more highly correlated than more distal CpGs. The rationale for choosing ζ_1 and ζ_2 is based on preliminary analyses that are provided in the Results Section.

To start, the cluster generating algorithm receives a sorted list of CpG base-positions from a given chromosome (i.e. chromosomal coordinates of CpGs) as input and outputs groups of clusters of CpGs, which we will refer to as super clusters. To achieve this goal, the algorithm starts with the smallest base-position and moves across the chromosome in ascending order of base-positions one CpG at a time. If a given CpG is less than $\zeta_1 = 3000$ base positions apart from its direct downstream neighbor, both CpGs will be assigned to the same cluster. If, on the other hand, the distance between these two neighbors exceeds the target threshold they will be assigned to separate clusters.

After all CpGs are assigned to clusters, a splitting step is performed in order to control the total number of CpGs per cluster. Each cluster containing more than $\Upsilon_1 = 10$ CpGs is split into two smaller, mutually exclusive clusters. This is achieved by calculating the base-pair distances between all direct neighbors in a cluster and putting them in descending order. Starting with the largest distance, the algorithm then checks whether splitting the cluster between the two respective neighboring CpGs will lead to at least one of the two new resulting clusters to contain less than or equal to 70% of CpGs from the original cluster. If this is the case, the cluster is split in two. If this is not the case, the algorithm keeps moving on to the next largest distance until the 70% stopping rule is satisfied and the splitting is performed. If any new cluster resulting from a split still contains more than Υ_1 CpGs, it and its descendants are repeatedly split in two, following the same steps outlined above, until all clusters contain sufficiently few CpGs.

The process that in-turn groups clusters into super clusters is directly analogous to how CpGs were grouped into clusters. First, clusters are put into ascending order according to the location of the genomic region that is spanned by their members. Next, the algorithm starts with the smallest location coordinate and moves across the chromosome one cluster at a time. If a given cluster region is less than $\zeta_2 = 30000$ base positions apart from the region of its direct downstream neighbor, both clusters will be assigned to the same super cluster. Neighboring clusters exceeding the target distance threshold are assigned to separate super clusters.

Finally, super clusters and its descendants containing more than Υ_2 clusters are repeatedly split into smaller super clusters, until all resulting super clusters contain $\Upsilon_2 = 8$ clusters or less. For each super cluster, base-pair distances between genomic cluster regions of direct neighbors are calculated and put into descending order. Starting with the two neighbors that are furthest apart, the algorithm will keep moving on to the next largest cluster separation distance until the two new resulting super clusters contain less than or equal to 70% of clusters from the original super cluster.

It is possible that in sparse genomic regions that contain very few CpGs this algorithm produces degenerate super clusters which each contain only a single cluster which itself contains only a single CpG. Whenever these cases occurred in this study, they were discarded, since the hier-

archical model would be overspecified and unsuitable for analysis. However, if our modelling approach was applied to a real study these special cases could still be separately analyzed utilizing the "TM" class of models, which do not require a structural hierarchy but are unable to leverage spatial correlation. For the practical purpose of faster processing speed of our simulations, another size restriction was also put into place. Even though we confirmed that 80 CpGs per super cluster were possible to process, super clusters that contained more than 44 CpGs were discarded without further analysis. This was deemed appropriate since the set of large super clusters containing 45 or more CpGs was relatively small (as reported in the Results Section).

4.5 Results

4.5.1 Preliminary Analyses

Plots of the autocorrelation of DNAm as a function of base-pair distance between CpGs revealed similar trends regardless of cell type and genomic location (Figure 4.2). While there was a difference between autosomes and chromosome X, the latter exhibiting higher frequencies of strongly correlated or anti-correlated CpGs, the general correlation trend with base pair distance was similar between autosomes and chromosome X (Figure C.1). Correlation values were overall very noisy, spanning the entire spectrum from -1 to 1, but tended to slightly concentrate towards positive values for distances of less than 50000 base pairs. Notably, a pronounced concentration towards 1.0 was observed among base-pair distances of less than 3000. In this range of close proximity, correlations tended to increase as base-pair distance decreased. Interpolating base-pair distance to 1, smoothed estimates of correlations yielded values in the range of 0.4 and 0.55. These properties motivated the utilized cluster grouping thresholds of $\zeta_1 = 3000$, aiming to capture close proximity correlation, and $\zeta_2 = 30000$, aiming to capture residual correlation in moderate proximity.

The threshold $\zeta_2 = 30000$ was also further motivated by another factor. Semi-degenerate super clusters that contain a very small number of total CpGs, are expected to struggle in adequately leveraging spatial correlation, since there is less information that can be borrowed from neighboring CpGs. As mentioned earlier, in the most extreme case of degenerate super clusters, which

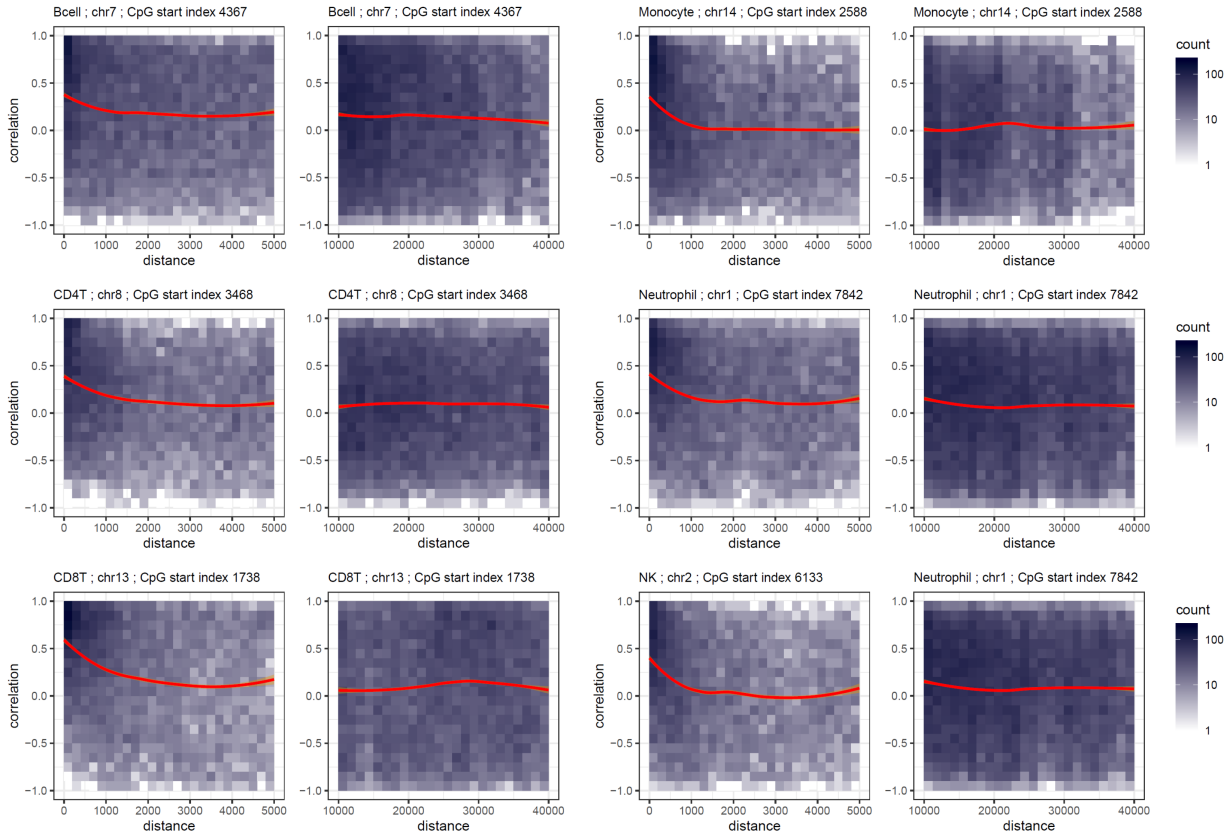


Figure 4.2: Pearson correlations of sample beta values for blood cell types as a function of base-pair distance. Orange lines in each plot represent loess smoothed correlation values via the "ggplot2" R-package. Similar patterns emerge regardless of cell type or genomic location. Correlation values exhibit higher concentration towards positive values. In the range of a base-pair distance of less than 3000, concentration towards higher correlation values appears to be more pronounced, a trend which diminishes as distance increases.

contain only a single CpG, fitting our hierarchical models is not even possible. Thus, ζ_2 was chosen such that the number CpGs that were assigned to degenerate and semi-degenerate super clusters was small. When choosing $\zeta_2 = 10000$, 2.4% of CpGs in the genome originated from degenerate super clusters, which was deemed too high. When choosing $\zeta_2 = 20000$, less than 1% of CpGs originated from degenerate super clusters, however, 1.6% of CpGs originated from super clusters containing only two or less CpGs. Choosing $\zeta_2 = 30000$ was appealing, since only 0.8% of CpGs originated from super clusters containing only two or less CpGs.

Lastly, the decision to focus on super clusters containing less than 45 CpGs in our simulation was based on the fact that most of the generated super clusters exhibited moderate sizes where

96% contained 44 CpGs or less and the largest observed super cluster exhibited a size of 66 CpGs.

4.5.2 Evaluating Marginal Model Fit

In simulation A, models were primarily compared based on their posterior mean RMSPE on a testing set, which was not utilized during the prediction of posterior distributions. Conducted simulation scenarios utilized datasets with a sample size of $N = 20$ and considered chromosome 1 and chromosome 13, as well as high, medium and low noise settings ($\sigma^{[B]}$ assuming 0.1, 0.05 and 0.01 respectively) for sampled beta values. After the runtime of two weeks expired, all simulation scenarios had processed 644 or more clusters, achieving a median number of 940 iterations.

As expected, prediction errors increased substantially with increasing standard deviation of the simulated beta values (Table 4.1). Across all simulated scenarios the choice of priors for variance parameters of beta values had a stronger effect on prediction error than the choice of model class: type 2 informative priors achieved the smallest, type 1 informative priors achieved the second smallest and weakly informative priors achieved the largest prediction errors. However, within each type of variance prior, spatial models consistently achieved smaller posterior mean RMSPE than non-spatial models. Prediction error differences between the three classes of spatial models when fixing the type of variance prior were small relative to each other, most scenarios differing by a value of less than 0.008. No substantial difference was observed between simulations based on chromosome 1 and those based on chromosome 13.

Since the value of p_D is only meaningful when comparing different models relative to each other for the same dataset, but not across different datasets, its values were first transformed. For each simulated dataset p_D values were divided by the p_D value of the TM model with weakly informative variance priors, yielding the complexity rates r_{pD}. In essence, these rates quantify by what factor a target model is more complex than the weakly informative version of the TM model. Similar to the prediction errors, average model complexity was more strongly affected by choice of prior distributions than the choice of model class and no substantial difference was observed between chromosome 1 and 13 (Table 4.2). Within each type of variance prior, average model

Table 4.1: Average posterior mean RMSPE of the testing data for a variety of different models in simulation type A. Each row in the table corresponds to a separate candidate model, while each column represents a different simulation scenario. Here, $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values. Row name suffixes denote the employed types of prior distribution: "wip" for weakly informative priors, "ip1" for informative priors of type 1 and "ip2" for informative priors of type 2. The type of employed prior distribution had a stronger effect on prediction error than the effect of model structure. Within each type of prior distribution, spatial models consistently achieve smaller average prediction errors than non-spatial models. Informative priors of type 2 consistently achieve the smallest prediction errors.

	$\sigma^{[B]}$=0.01 chr1	$\sigma^{[B]}$=0.01 chr13	$\sigma^{[B]}$=0.05 chr1	$\sigma^{[B]}$=0.05 chr13	$\sigma^{[B]}$=0.1 chr1	$\sigma^{[B]}$=0.1 chr13
SCM1.ip2	0.026	0.028	0.092	0.092	0.176	0.175
SCM2.ip2	0.027	0.028	0.092	0.092	0.176	0.176
SCM3.ip2	0.026	0.028	0.092	0.092	0.175	0.175
TM.ip2	0.028	0.028	0.096	0.096	0.189	0.190
SCM1.ip1	0.052	0.055	0.098	0.099	0.177	0.176
SCM2.ip1	0.052	0.055	0.098	0.100	0.177	0.178
SCM3.ip1	0.052	0.055	0.098	0.099	0.176	0.177
TM.ip1	0.054	0.057	0.103	0.104	0.191	0.191
SCM1.wip	0.082	0.086	0.127	0.129	0.207	0.208
SCM2.wip	0.082	0.086	0.127	0.130	0.208	0.210
SCM3.wip	0.082	0.086	0.127	0.130	0.208	0.210
TM.wip	0.087	0.092	0.139	0.142	0.234	0.236

Table 4.2: Average model complexity rates for a variety of different models in simulation type A. Each row in the table corresponds to a separate model, while each column represents a different simulation scenario. $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values. Row name suffixes denote the employed types of prior distribution: "wip" for weakly informative priors, "ip1" for informative priors of type 1 and "ip2" for informative priors of type 2. Values of complexity rates denote by what factor on average a target model is more complex than the "TM.wip" model. Within each type of prior distribution, spatial models consistently achieved lower average complexity than non-spatial models.

	$\sigma^{[B]}$=0.01	$\sigma^{[B]}$=0.01	$\sigma^{[B]}$=0.05	$\sigma^{[B]}$=0.05	$\sigma^{[B]}$=0.1	$\sigma^{[B]}$=0.1
	chr1	chr13	chr1	chr13	chr1	chr13
SCM1.ip2	0.95	0.94	0.56	0.59	0.34	0.35
SCM2.ip2	0.95	0.94	0.56	0.59	0.34	0.36
SCM3.ip2	0.95	0.94	0.57	0.59	0.35	0.36
TM.ip2	0.96	0.94	0.59	0.62	0.41	0.42
SCM1.ip1	0.93	0.91	0.84	0.82	0.51	0.52
SCM2.ip1	0.93	0.91	0.84	0.82	0.51	0.53
SCM3.ip1	0.93	0.91	0.84	0.82	0.51	0.52
TM.ip1	0.94	0.92	0.86	0.84	0.56	0.57
SCM1.wip	0.99	0.98	0.98	0.98	0.98	0.97
SCM2.wip	0.99	0.98	0.98	0.98	0.98	0.97
SCM3.wip	0.99	0.98	0.99	0.98	0.98	0.97
TM.wip	1	1	1	1	1	1

complexity in spatial models was consistently lower than the TM model. In low-noise settings, all complexity rates assumed values between 0.91 and 1.0, suggesting only small differences in model complexity, with informative type 1 priors achieving lowest complexity among priors, followed by informative type 2 priors. In settings with larger noise, informative type 2 priors achieved the smallest and informative priors of type 1 achieved the second smallest complexity rates. Average model complexity rates decreased with increasing noise levels, achieving a lowest value of 0.56 for $\sigma^{[B]} = 0.05$ and a lowest value of 0.34 for $\sigma^{[B]} = 0.1$.

Considering that all spatial model types achieved comparable performance measures, model type SCM2 was chosen for further evaluation. The rationale behind picking this candidate was that, in contrast to SCM1, it did not make assumptions about relationships between cell types and, in contrast to SCM3, it structurally honored the way CpGs were incorporated into super clusters.

4.5.3 Performance in the Two-Arm Design When Cell Proportions Are Balanced

Simulation B utilized datasets with a sample size of $N = 40$ where 20 samples each belonged to the two treatment arms, was restricted to chromosome 1 and considered the same high, medium and low noise settings for $\sigma^{[B]}$ as in simulation A. Different levels of effect size variation were captured by considering effect size standard deviations ($\sigma^{[C]}$) of 0.01, 0.05 and 0.1, as well as effect size correlations ($\rho^{[C]}$) of 0.5 and 0.2. After the runtime of two weeks expired, all simulation scenarios had processed 407 or more clusters, achieving a median number of 419 iterations.

In the alternative space ($\Delta = 0.2$), average posterior mean RMSPEs behaved analogously to simulation A, being most strongly affected by an increasing trend with noise level ($\sigma^{[B]}$) while also decreasing from weakly informative to informative type 1 priors and further from informative type 1 to informative type 2 priors (Table 4.3). Within any given prior category the spatial SCM2 model consistently achieved lower prediction errors than the non-spatial TM model. Within each simulation scenario, the smallest overall value of average mean RMSPE was consistently observed for the SCM model with informative priors of type 2.

Again, analogously to simulation A, average model complexity in the alternative space was more strongly affected by choice of prior distributions than the choice of candidate model (Table 4.4) and within each type of variance prior, average complexity rates were consistently lower in SCM2 compared to TM. Complexity rates also generally decreased with increasing noise levels reaching their lowest values when $\sigma^{[B]} = 0.1$. SCM2 models achieved the smallest overall model complexity in high and medium noise settings, as well as second smallest overall model complexity in the low noise setting.

Since SCM2 with informative type 2 priors was both among models exhibiting the best fit characteristics in simulation A and also the best candidate model in simulation B, statistical power was compared between SCM2 and TM models, using informative priors of type 2.

Testing for a difference in mean methylation between two study arms was first evaluated via a classical, fixed posterior quantile decision rule: Reject H_0, i.e. there being no difference in mean methylation for a specific CpG, if the 95% credible interval of the posterior mean difference for said

Table 4.3: Average posterior mean RMSPE of the testing data for the TM and SCM2 candidate models in simulation type B when $\Delta = 0.2$. Errors were all based on the evaluation of chromosome 1. Each row in the table corresponds to a separate candidate model, while each column represents a different simulation scenario. $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values and $\sigma^{[C]}$ denotes the effect size standard deviation. Since there was no pronounced difference between effect size correlations ($\rho^{[C]}$) of 0.5 and 0.2 in any of the considered scenarios, only results for $\rho^{[C]} = 0.5$ are shown. Row name suffixes denote the employed types of prior distribution: "wip" for weakly informative priors, "ip1" for informative priors of type 1 and "ip2" for informative priors of type 2. The type of employed prior distribution had a stronger effect on prediction error than the effect of model structure. Within each type of prior distribution, spatial SCM2 models consistently achieved smaller average prediction errors than non-spatial models. Informative priors of type 2 consistently achieve the smallest prediction errors.

$\sigma^{[B]}=\ldots$	0.01	0.01	0.01	0.05	0.05	0.05	0.1	0.1	0.1
$\sigma^{[C]}=\ldots$	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
SCM2.ip2	0.022	0.022	0.023	0.089	0.089	0.090	0.172	0.172	0.173
TM.ip2	0.024	0.024	0.024	0.097	0.097	0.097	0.191	0.191	0.191
SCM2.ip1	0.040	0.040	0.041	0.092	0.092	0.093	0.173	0.173	0.174
TM.ip1	0.044	0.044	0.044	0.100	0.100	0.100	0.192	0.192	0.192
SCM2.wip	0.049	0.049	0.050	0.102	0.102	0.103	0.185	0.185	0.186
TM.wip	0.056	0.056	0.056	0.115	0.115	0.115	0.211	0.211	0.211

Table 4.4: Average model complexity rates for a variety of TM and SCM2 candidate models in simulation type B when $\Delta = 0.2$. Rates were all based on the evaluation of chromosome 1. Each row in the table corresponds to a separate model, while each column represents a different simulation scenario. $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values and $\sigma^{[C]}$ denotes the effect size standard deviation. Since there was no pronounced difference between effect size correlations ($\rho^{[C]}$) of 0.5 and 0.2 in any of the considered scenarios, only results for $\rho^{[C]} = 0.5$ are shown. Row name suffixes denote the employed types of prior distribution: "wip" for weakly informative priors, "ip1" for informative priors of type 1 and "ip2" for informative priors of type 2. Values of complexity rates denote by what factor on average a target model is more complex than the "TM.wip" model. Within each type of prior distribution, spatial SCM2 models consistently achieve lower average complexity than non-spatial models.

$\sigma^{[B]}=\ldots$	0.01	0.01	0.01	0.05	0.05	0.05	0.1	0.1	0.1
$\sigma^{[C]}=\ldots$	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
SCM2.ip2	0.96	0.97	0.97	0.50	0.50	0.50	0.29	0.29	0.29
TM.ip2	0.97	0.97	0.97	0.56	0.56	0.56	0.40	0.40	0.40
SCM2.ip1	0.95	0.95	0.95	0.85	0.85	0.84	0.48	0.48	0.47
TM.ip1	0.97	0.97	0.97	0.87	0.87	0.87	0.55	0.55	0.55
SCM2.wip	0.98	0.98	0.99	0.98	0.98	0.98	0.97	0.97	0.97
TM.wip	1	1	1	1	1	1	1	1	1

CpG does not include 0. Employing this decision rule led to type 1 error rates varying drastically based on the simulation scenario (Table 4.5). For a fixed value of $\sigma^{[C]}$ and a given model, type 1 error rates consistently increased with beta value noise levels $\sigma^{[B]}$. Similarly, for a fixed value of $\sigma^{[B]}$ and a given model, type 1 error rates consistently increased with $\sigma^{[C]}$. For a fixed $\sigma^{[B]}$, a fixed $\sigma^{[C]}$ and a given model, type 1 error differences between $\rho^{[B]} = 0.5$ and $\rho^{[B]} = 0.2$ were small, never exceeding 0.03. Generally, the $\rho^{[B]} = 0.5$ setting appeared to achieve slightly smaller error rates than the $\rho^{[B]} = 0.2$ setting for TM models and slightly larger error rates for SCM2 models, though this trend was not consistent. Lowest type 1 error rates, 0.01 for the TM model and 0.04 for the SCM2 model , were achieved when both beta value noise level and effect size standard deviation were low, i.e. $\sigma^{[B]} = \sigma^{[C]} = 0.01$. On the other hand, the highest type 1 error rates, 0.36 for the TM model and 0.39 for the SCM2 model, were achieved when beta value noise levels were low $\sigma^{[B]} = 0.01$ and effect size standard deviation was large $\sigma^{[C]} = 0.1$. In all other scenarios, type 1 errors were more consistent, ranging from 0.07 to 0.13 in the TM model and ranging from 0.15 to 0.26 in the SCM2 model. Generally, type 1 errors were inflated in SCM2 models, on average being 2.5 times higher than rates obtained from TM models.

In the alternative hypothesis space, beta value noise exhibited the strongest effect, overall statistical power decreasing with $\sigma^{[B]}$. For TM models, power was ranging from 0.82 to 0.97 when $\sigma^{[B]} = 0.01$, from 0.23 to 0.26 when $\sigma^{[B]} = 0.05$ and from 0.14 to 0.15 when $\sigma^{[B]} = 0.01$. For SCM2 models, power was ranging from 0.89 to 1.0 when $\sigma^{[B]} = 0.01$, from 0.53 to 0.57 when $\sigma^{[B]} = 0.05$ and from 0.37 to 0.38 when $\sigma^{[B]} = 0.01$. Even though SCM2 power for the fixed quantile decision rule was substantially larger than TM power in all cases, a direct comparison of the two models is confounded by the inflation of type 1 error rates in SCM models. To account for this, a modified rejection rule was considered next, which calibrated type 1 error rates in each simulation scenario.

For each simulation scenario, the calibrated rejection rule was formulated as follows: Among all credible intervals, collected during the simulation, first pick the $q_1\%$ credible interval of the SCM2 model, such that its corresponding type 1 error rate e_1 is the largest but also satisfies $e_1 <$

Table 4.5: Overall power analysis comparing spatial SCM2 and non-spatial TM models when testing for differential methylation of individual CpGs in simulation B for the following rejection rule: Reject H_0 if the 95% credible interval of the mean difference between the two arms excludes 0. "iterations" refers to the number simulations used to estimate operating characteristics. "t1e" denotes type 1 error rate and "pow" denotes statistical power. $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values and $\sigma^{[C]}$ denotes the effect size standard deviation.

$\sigma^{[B]}$	$\sigma^{[C]}$	$\rho^{[C]}$	iterations	t1e.TM	t1e.SCM2	$\frac{t1e.SCM2}{t1e.TM}$	pow.TM	pow.SCM2
0.01	0.01	0.5	413	0.01	0.04	2.98	0.97	>0.99
0.01	0.01	0.2	414	0.01	0.03	3.01	0.97	>0.99
0.01	0.05	0.5	422	0.13	0.17	1.37	0.93	0.98
0.01	0.05	0.2	414	0.13	0.15	1.15	0.93	0.99
0.01	0.1	0.5	422	0.36	0.39	1.09	0.82	0.89
0.01	0.1	0.2	416	0.36	0.37	1.01	0.82	0.89
0.05	0.01	0.5	420	0.07	0.20	2.90	0.23	0.57
0.05	0.01	0.2	411	0.07	0.20	2.88	0.24	0.57
0.05	0.05	0.5	414	0.08	0.21	2.74	0.24	0.56
0.05	0.05	0.2	410	0.08	0.20	2.58	0.25	0.56
0.05	0.1	0.5	415	0.11	0.24	2.24	0.26	0.54
0.05	0.1	0.2	411	0.11	0.21	1.93	0.26	0.53
0.1	0.01	0.5	413	0.09	0.25	2.70	0.14	0.38
0.1	0.01	0.2	410	0.09	0.25	2.65	0.14	0.38
0.1	0.05	0.5	413	0.10	0.25	2.58	0.14	0.38
0.1	0.05	0.2	410	0.10	0.25	2.55	0.14	0.38
0.1	0.1	0.5	420	0.11	0.26	2.48	0.15	0.38
0.1	0.1	0.2	407	0.11	0.25	2.35	0.15	0.37

0.1. If none of the collected intervals leads to an error rate of less than 0.1, then q_1 is instead chosen to minimize type 1 error. After q_1 is chosen, identify the $q_2\%$ credible interval of the TM model such that its corresponding type 1 error rate e_2 is the smallest but also satisfies $e_2 > e_1$. In summary, this rejection rule simultaneously accomplishes two goals. Firstly, it consistently calibrates type 1 error, in both TM and SCM2, to either 10% or, if not possible, to the smallest observed error rate among collected quantiles. Secondly, it always leads to TM models having slightly higher type 1 error rates than in SCM2 models. This second goal provides a small advantage to TM models when comparing statistical power, since statistical power generally increases as a function of type 1 error.

Calibration of $e_1 \approx e_2 \approx 0.1$ was approximately achieved in all simulation scenarios, except for two cases (Table 4.6). More specifically, when $\sigma^{[B]} = 0.01$ and $\sigma^{[C]} = 0.1$ type 1 errors were calibrated to approximately 0.17 for $\rho^C = 0.5$ and to approximately 0.14 for $\rho^C = 0.2$. After calibration, a similar, strong, decreasing effect of $\sigma^{[B]}$ on statistical power was observed. For TM models, power was ranging from 0.62 to 1.0 when $\sigma^{[B]} = 0.01$, from 0.24 to 0.29 when $\sigma^{[B]} = 0.05$ and from 0.14 to 0.15 when $\sigma^{[B]} = 0.01$. For SCM2 models, power was ranging from 0.74 to 1.0 when $\sigma^{[B]} = 0.01$, from 0.32 to 0.39 when $\sigma^{[B]} = 0.05$ and from 0.16 to 0.17 when $\sigma^{[B]} = 0.01$. After calibration the SCM2 model still yielded consistently more powerful tests than the TM model. The improvement in power was particularly striking in the moderate noise level scenarios ($\sigma^{[B]} = 0.05$) in which an approximate 10% increase in power was observed. The consistent improvement in power of SCM2 over TM was also observed when calibrating simulation scenarios to lower type 1 error rates, such as 5%, though the magnitude of the difference was smaller. However, since with the currently collected credible intervals (the widest credible interval covering 99.9% of the posterior distribution) less than half of the considered scenarios could be calibrated to these smaller error rates, these results were not included into this section.

Statistical power was also evaluated as a function of cluster size. Neither the number of clusters within a super cluster nor the total base-pair distance spanned by a super cluster appeared to have an effect on type 1 error or power. However, the total number of CpGs contained within a super

Table 4.6: Overall power analysis comparing spatial SCM2 and non-spatial TM models when testing for differential methylation of individual CpGs in simulation B when type 1 error is calibrated: In each scenario, the rejection rule employs the $q\%$ credible interval, where q is chosen such that the type 1 error is controlled at a 10% level. In cases where the widest considered credible interval leads to a larger type 1 error rate in spatial models, both model classes are calibrated to this error rate instead. "iterations" refers to the number simulations used to estimate operating characteristics. "t1e" denotes type 1 error rate and "pow" denotes statistical power. $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values and $\sigma^{[C]}$ denotes the effect size standard deviation.

$\sigma^{[B]}$	$\sigma^{[C]}$	$\rho^{[C]}$	iterations	t1e.TM	t1e.SCM2	$\frac{t1e.SCM2}{t1e.TM}$	pow.TM	pow.SCM2
0.01	0.01	0.5	413	0.10	0.10	1.00	>0.99	>0.99
0.01	0.01	0.2	414	0.10	0.10	1.00	>0.99	>0.99
0.01	0.05	0.5	422	0.10	0.10	0.99	0.91	0.96
0.01	0.05	0.2	414	0.10	0.10	0.98	0.91	0.98
0.01	0.1	0.5	422	0.17	0.17	0.96	0.66	0.75
0.01	0.1	0.2	416	0.14	0.14	0.95	0.62	0.74
0.05	0.01	0.5	420	0.10	0.10	0.99	0.29	0.39
0.05	0.01	0.2	411	0.10	0.10	1.00	0.29	0.39
0.05	0.05	0.5	414	0.10	0.10	0.99	0.28	0.37
0.05	0.05	0.2	410	0.10	0.10	1.00	0.28	0.38
0.05	0.1	0.5	415	0.10	0.09	0.98	0.24	0.32
0.05	0.1	0.2	411	0.10	0.10	1.00	0.25	0.35
0.1	0.01	0.5	413	0.10	0.10	0.99	0.15	0.17
0.1	0.01	0.2	410	0.10	0.10	0.99	0.14	0.17
0.1	0.05	0.5	413	0.10	0.10	0.99	0.14	0.17
0.1	0.05	0.2	410	0.10	0.10	1.00	0.14	0.17
0.1	0.1	0.5	420	0.11	0.10	0.99	0.15	0.17
0.1	0.1	0.2	407	0.10	0.10	0.99	0.14	0.16

cluster did appear to affect the type 1 error inflation of SCM2 models. When utilizing the decision rule based on 95% credible intervals of mean difference in methylation, the average of type 1 error inflation factors of SCM2 (i.e. the factor by which type 1 error of SCM2 is larger than type 1 error of TM) across all simulated scenarios decreased as the total number of CpGs per super cluster (or short TNCPS) decreased. Inflation factors (as previously mentioned) averaged 2.5 when considering all super clusters, 2.1 when TNCPS were less than 30, 1.6 when TNCPS were less than 20, 1.3 when TNCPS were less than 15, and 1.2 when TNCPS were less than 10. Smaller super cluster sizes were not evaluated since too few simulations of such super clusters were generated (only 43 clusters exhibited TNCPS of less than 10). This change was caused by type 1 error rates decreasing for SCM2 models while type 1 error rates for TM models remained stable, only showing small random deviations when varying maximum TNCPS.

4.5.4 Performance in the Two-Arm Design When Cell Proportions Are Unbalanced

Simulation C considered the same settings in term of sample sizes in each treatment arm and target chromosome, but only considered high and low levels, i.e. values of 0.1 and 0.01 respectively, for both $\sigma^{[B]}$ and $\sigma^{[C]}$, and fixed $\rho^{[C]}$ to a value of 0.5. After the runtime of one week expired, all simulation scenarios had processed 232 or more clusters, achieving a median number of 236 iterations.

For the overall differential methylation analysis across all cell-types, employing the 95% credible interval decision rule led to similar type 1 error rates compared to simulation B, deviating by less than 0.04 in all cases for both spatial and non-spatial models. (Table 4.7). General trends observed in simulation B also applied in simulation C: for fixed $\sigma^{[C]}$ type 1 error increased with $\sigma^{[B]}$ and power decreased with $\sigma^{[B]}$; for fixed $\sigma^{[B]}$ type 1 error decreased with $\sigma^{[C]}$ and power tended to decrease with $\sigma^{[C]}$. A notable difference was a decrease in type 1 error inflation of spatial models compared to non-spatial models in high noise level scenarios ($\sigma^{[B]} = 0.1$). This was caused by a 0.03 increase of type 1 error compared to simulation B for TM models while type 1 error for SCM2 models slightly decreased compared to simulation B. Overall power for this decision rule

Table 4.7: Overall power analysis comparing spatial SCM2 and non-spatial TM models when testing for differential methylation of individual CpGs in simulation C. The "rule" column specifies the employed rejection rule. Rule "F" rejects H_0 if the 95% credible interval of the mean difference between the two arms excludes 0. In rule "C" type 1 error is calibrated such that in each scenario the rejection rule employs the q% credible interval, where q is chosen such that the type 1 error is controlled at a 10% level. In cases where the widest considered credible interval leads to a larger type 1 error rate in spatial models, both model classes are calibrated to this error rate instead. "iterations" refers to the number simulations used to estimate operating characteristics. "t1e" denotes type 1 error rate and "pow" denotes statistical power. $\sigma^{[B]}$ denotes the standard deviation employed when drawing beta values and $\sigma^{[C]}$ denotes the effect size standard deviation.

rule	$\sigma^{[B]}$	$\sigma^{[C]}$	iterations	t1e.TM	t1e.SCM2	$\frac{t1e.SCM2}{t1e.TM}$	pow.TM	pow.SCM2
F	0.01	0.01	232	0.02	0.06	2.97	0.72	0.89
F	0.01	0.1	236	0.32	0.38	1.16	0.66	0.80
F	0.1	0.01	236	0.13	0.23	1.84	0.19	0.38
F	0.1	0.1	236	0.14	0.25	1.79	0.21	0.38
C	0.01	0.01	232	0.10	0.10	1.00	0.83	0.92
C	0.01	0.1	236	0.20	0.20	1.00	0.53	0.63
C	0.1	0.01	236	0.10	0.10	1.00	0.15	0.20
C	0.1	0.1	236	0.10	0.09	0.98	0.15	0.18

was decreased in low noise level scenarios ($\sigma^{[B]} = 0.01$) compared to simulation B, decreasing by 0.09 or more in all cases for both spatial and non-spatial models. On the other hand, in high level noise scenarios for both types of models overall power of the 95% credible interval decision rule increased compared to simulation B, though to a smaller degree.

Employing rejection rules that calibrated type 1 error to 10% in the overall analysis, led to a decreasing in statistical power compared to simulation B in low noise level scenarios for both spatial and non-spatial models. On the other hand, statistical power for testing difference in mean methylation in high noise level scenarios was very similar to simulation B for TM models, yet slightly increased compared to simulation B in SCM2 models. Statistical power also remained consistently higher in SCM2, achieving improvements over TM in the range of 0.04 to 0.09.

In the cell type specific analysis of difference in mean methylation between the two study arms type 1 error rates of cell types when employing the 95% credible interval decision rule tended to scatter around similar scenario averages within each model type, i.e. within TM or SCM2 (Figure 4.3). Type 1 errors were generally highest when both $\sigma^{[B]} = 0.01$ and $\sigma^{[C]} = 0.1$, and generally

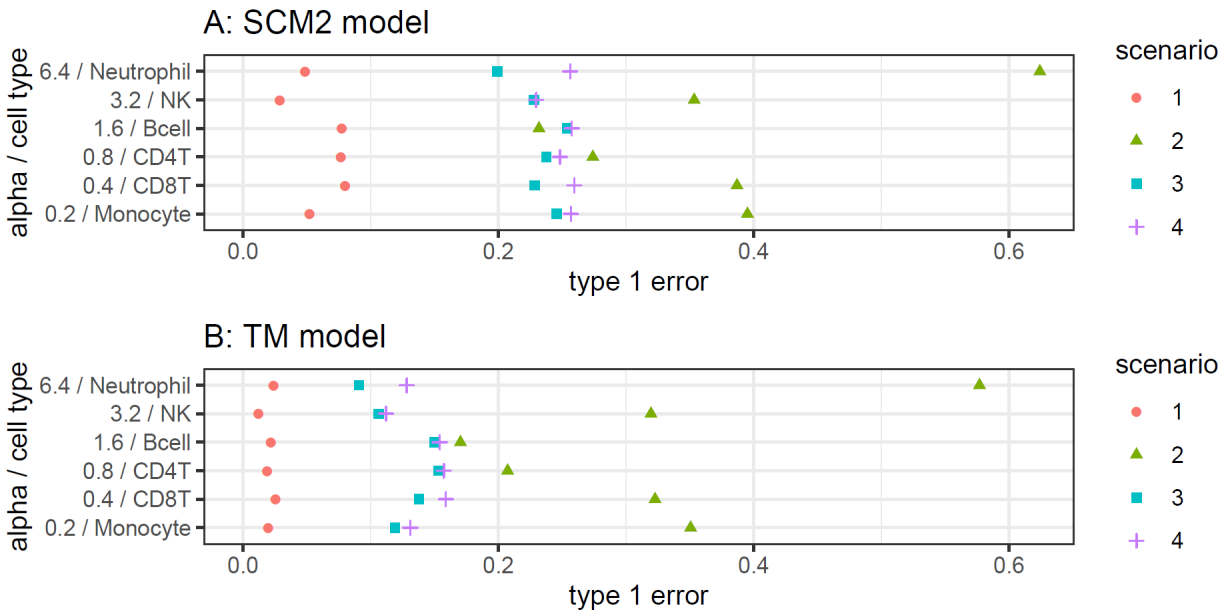


Figure 4.3: Cell type specific analysis comparing type 1 error for spatial SCM2 and non-spatial TM models when testing for differential methylation of individual CpGs in simulation C using the fixed, 95% credible interval decision rule. Here, "alpha" refers to the cell type specific parameter that was used to draw cell proportions via a Dirichlet distribution; a higher value corresponds to consistently drawing higher cell proportions. "scenario" refers to the following configurations: $1 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.01, 0.01)$; $2 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.01, 0.1)$; $3 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.1, 0.01)$; $4 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.1, 0.1)$. For each model type scenarios tend to on average produce similar error rates. No clear trend with cell proportion is observed.

lowest when both $\sigma^{[B]} = 0.01$ and $\sigma^{[C]} = 0.01$. However, within each simulation scenario the degree to which type 1 errors of an individual cell type deviated from a common average of other cell types did not appear to follow any clear trend. A specifically strong deviation from other cell types was observed for both TM and SCM2 models in the $\sigma^{[B]} = 0.01, \sigma^{[C]} = 0.1$ scenario, in which Neutrophils exhibited a type 1 error rate that was 1.5 times higher than the error rate for any other cell type. While type 1 errors remained consistently inflated in SCM2 models compared to TM models, the degree of inflation also varied unsystematically across cell types and simulation scenarios. Overall, no clear trend of type 1 error with magnitude of cell proportion was observed.

After type 1 errors were calibrated to 10%, cell type specific, statistical power followed a similar trend for each cell type, independent of the target candidate model: decreasing substantially when increasing $\sigma^{[B]}$ from 0.01 to 0.1, but also decreasing slightly when increasing $\sigma^{[C]}$ from 0.01

to 0.1 for any fixed value of $\sigma^{[B]}$ (Figure 4.4). Neutrophils, which were consistently sampled with the largest cell proportions, exhibited the largest statistical power, clearly separating from all other cell types for both spatial and non-spatial models. However, across the remaining other five cell types statistical power did not appear to follow any clear trend with magnitude of cell proportion, though there was a tendency for Bcell and NK cells to perform consistently worse than other cell types. Analogously to results from the overall analysis of mean methylation, statistical power in any cell type was consistently larger for spatial SCM2 models compared to non-spatial TM models. Excluding the low noise scenario in which both models achieved a power of approximately 1.0, the increase in statistical power gained by utilizing spatial models ranged from 0.01 to 0.15, achieving a median value of 0.06. The SCM2 model appeared to provide the largest increase in power primarily in those scenarios, in which power of the TM model was in the range between 0.2 and 0.8; a condition which did not appear to be associated with any specific cell type. A detailed summary of cell type specific power when calibrating type 1 error is provided in Table C.1.

4.6 Discussion

The empirical correlation plots of CpG methylation observed in this study exhibit a smoothed trend very similar to one previously observed by Liu et al. (2014), who investigated whole-blood samples from 247 healthy individuals. In their study, correlation of CpGs for a base-pair distance of 1 also interpolated to values close to 0.4 and correlation also decreased with base-pair distance in an exponential decay shape tending either towards 0 or a small positive value. Notably, most of the observed decay in correlation was also already completed at a base-pair distance of 3000, with very little change beyond larger distances. Even though the researchers evaluated methylation from bulk samples, this observation is likely consistent with our findings that all isolated cell types exhibited a similar pattern. To elaborate this point, consider a simplified example in which methylation of different cell types is independent and methylation within the same cell type follows the same spatial correlation structure, regardless of cell-type. We assume that for any given sample i, for any two CpGs u, v and for any two isolated cell types h, k correlation in methylation levels can be

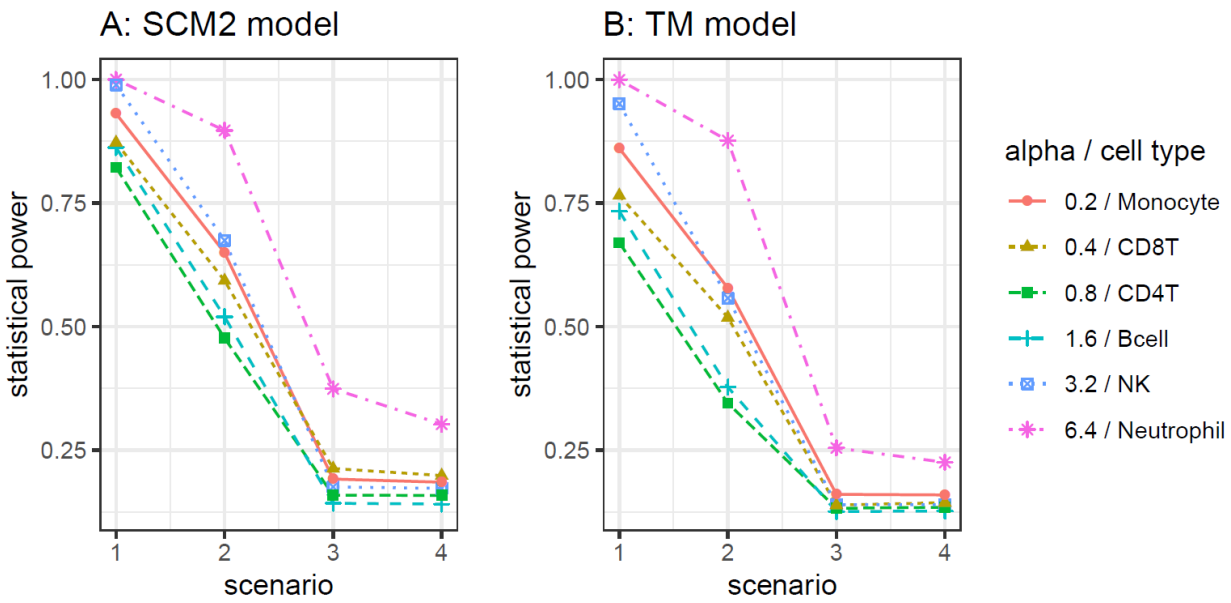


Figure 4.4: Cell type specific analysis comparing statistical power for spatial SCM2 and non-spatial TM models when testing for differential methylation of individual CpGs in simulation C when type 1 error is calibrated: In each scenario, the rejection rule employs the $q\%$ credible interval, where q is chosen such that the type 1 error is controlled at a 10% level. In cases where the widest considered credible interval leads to a larger type 1 error rate in spatial models, both model classes are calibrated to this error rate instead. Here, "alpha" refers to the cell type specific parameter that was used to draw cell proportions via a Dirichlet distribution; a higher value corresponds to consistently drawing higher cell proportions. "scenario" refers to the following configurations: $1 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.01, 0.01)$; $2 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.01, 0.1)$; $3 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.1, 0.01)$; $4 \rightarrow (\sigma^{[B]}, \sigma^{[C]}) = (0.1, 0.1)$. Statistical power follows a similar trend across scenarios for each model type and for each cell type. No clear trend with cell proportion is observed. For each cell type, spatial models achieve consistently higher power than non-spatial models.

expressed as: $Corr(Z_{ihu}, Z_{ikv}) = I[h = k] \cdot \rho(u,v)$ where $\rho(u,v)$ denotes the correlation function between u, v and $I[h = k]$ is an indicator function returning 1 if $h = k$ and 0 otherwise. Further let $Var(Z_{ihu}) = \psi^2_{hu}$, let $X_{hu} = \sum_{h=1}^{H} w_{ih} Z_{ihu}$ and assume that the cell proportions w_{ih} are known, fixed quantities. We can then deduce the following:

$$
\begin{aligned}
Cov(X_{iu}, X_{iv}) &= Cov\left(\sum_{h=1}^{H} w_{ih} Z_{ihu}\,,\, \sum_{h=1}^{H} w_{ih} Z_{ihv}\right) = \sum_{h=1}^{H}\sum_{g=1}^{H} Cov\left(w_{ih} Z_{ihu}\,,\, w_{ig} Z_{igv}\right) && (4.1)\\
&= \sum_{h=1}^{H}\sum_{g=1}^{H} w_{ih} w_{ig} \cdot Cov\left(Z_{ihu}\,,\, Z_{igv}\right) && (4.2)\\
&= \sum_{h=1}^{H}\sum_{g=1}^{H} w_{ih} w_{ig} \cdot \psi_{hu}\psi_{gv} \cdot Corr\left(Z_{ihu}\,,\, Z_{igv}\right) && (4.3)\\
&= \sum_{h=1}^{H}\sum_{g=1}^{H} w_{ih} w_{ig} \cdot \psi_{hu}\psi_{gv} \cdot I[h = g]\rho(u,v) && (4.4)\\
&= \rho(u,v) \cdot \sum_{h=1}^{H} \sum_{g:g=h} I[h = g] \cdot w_{ih} w_{ig} \cdot \psi_{hu}\psi_{gv} && (4.5)\\
&\quad + \rho(u,v) \cdot \sum_{h=1}^{H} \sum_{g:g\neq h} I[h = g] \cdot w_{ih} w_{ig} \cdot \psi_{hu}\psi_{gv} && (4.6)\\
&= \rho(u,v) \cdot \sum_{h=1}^{H} I[h = h] \cdot w_{ih} w_{ih} \cdot \psi_{hu}\psi_{hv} + 0 && (4.7)\\
&= \rho(u,v) \cdot \sum_{h=1}^{H} w^2_{ih} \cdot \psi_{hu}\psi_{hv} && (4.8)\\
\Rightarrow Var(X_{iu}) &= Cov(X_{iu}, X_{iu}) = \rho(u,u) \cdot \sum_{h=1}^{H} w^2_{ih} \cdot \psi_{hu}\psi_{hu} = \sum_{h=1}^{H} w^2_{ih} \cdot \psi^2_{hu} && (4.9)\\
\Rightarrow Corr(X_{iu}, X_{iv}) &= \rho(u,v) \cdot \frac{\sum_{h=1}^{H} w^2_{ih}\psi_{hu}\psi_{hv}}{\sqrt{\sum_{m=1}^{H} w^2_{im}\psi^2_{mu}}\sqrt{\sum_{o=1}^{H} w^2_{io}\psi^2_{ov}}} && (4.10)
\end{aligned}
$$

This means that as long as within each cell type variability of methylation is relatively stable among proximal CpGs, i.e. $\psi^2_{hu} \approx \psi^2_{hv}$ holds roughly for all h, then $Corr(X_{iu}, X_{iv}) \approx \rho(u,v)$ and the correlation function of bulk methylation X will follow a similar shape as the correlation function of cell type specific methylation Z.

This insight suggests that cell specific variances may be either the same or at least themselves

correlated among proximal CpGs and also that the findings in this study could be potentially more generally applicable to whole blood methylation samples of healthy individuals. However, care should be taken when considering other sample populations or extensions of the model. The types of cells, their function and composition will vary drastically from tissue to tissue. It thus remains questionable whether decay in correlation with base-pair distance will be as similar between other cell types in other tissues. Even when solely considering samples from whole blood, substituting $\psi_{hj} = \psi_h$ in the likelihood is not necessarily guaranteed to improve performance of the models and may even depend on the size of a given CpG cluster. Evaluations of more datasets and different tissues will be necessary to explore these relationships in the future.

High noise levels in the empirical correlation plots provide justification for the cluster forming and hierarchical shrinking approach employed in this study. At first glance, the smooth loess fit of correlation values may tempt the observer to employ distance-based correlation structures such as autoregressive or Markov structures when fitting the model. However, the high noise levels suggest that these approaches may not be appropriate. Indeed, preliminary tests when fitting these types of models to the data resulted in inferior model fit and poor convergence behavior, though no definitive systematic comparison was conducted to follow-up on this observation. However, if we assume that instead proximal CpGs correlate within clusters and that proximal clusters themselves correlate within a super cluster, then it would be reasonable to expect that the average effect of these hierarchies would result in correlation decreasing with base-pair distance. At the same time, it also makes sense that individual pairs of CpGs would be able to highly deviate from this average trend. Lastly, fitting separate models to each super-cluster allows the degree to which CpGs correlate within clusters and to which clusters correlate with each other to change based on their genomic location.

It should be noted that in biological reality relationships are likely more complex than the simple hierarchy proposed in this study. Indeed, different or more complex structures may be discovered in the future that lead to more powerful tests and more robust estimation. Currently clusters of biologically related CpGs are identified solely based on base-pair distance between CpGs, but

many CpGs are already known to either be co-methylated with other CpGs or to belong to the same functional group as other CpGs. Public databases already provide annotations which, for example, group CpGs into so called “island”, “shores” and “shelfs”, which are often associated with specific genes. Forming clusters and their hierarchies based on these annotations could potentially do a better job of grouping closely related CpGs and improve performance. Even if we assume the current cluster forming strategy is superior to annotation-based approaches, there are more opportunities for improvement. There is no guarantee that the employed cluster grouping parameters $\zeta_1, \zeta_2, \Upsilon_1, \Upsilon_2$ will form the best or most biologically sound structural hierarchies. However, the current configuration appears to at least be reasonable, based on the observed improvements in performance over the non-spatial models.

The fact that in simulation A there is no pronounced difference in performance between SCM1, SCM2 and SCM3, could potentially suggest that simpler hierarchies may be sufficient to model spatial correlation. If shrinking different cluster means towards the overall super cluster mean is indeed as efficient as directly shrinking all CpGs in a super cluster towards the overall super cluster mean, then it might be more beneficial to simply form larger clusters without hierarchies and fit a separate model to each cluster. However, it is not necessarily obvious whether such a simplified approach would lead to better statistical power, since the super cluster structure may also be beneficial for stabilizing estimates of variance components and since different configurations of prior distributions and cluster grouping parameters may or may not lead to a more pronounced difference between candidate models. In the future, a systematic evaluation of how the interplay between cluster forming strategies, candidate models and prior configurations affects estimation and statistical inference is needed.

The observation that in simulation B and C type 1 error rates of SCM2 models were consistently inflated compared to TM models when utilizing fixed posterior credible intervals is noteworthy. The fact that in simulation B the rate of inflation decreased as the total number of CpGs per super cluster decreased, could potentially suggest that either effects of spatial correlation are diluted in large clusters or that as the number of CpGs increases models tend to become overly

confident about estimates of mean methylation. Both of these problems could potentially be addressed by refining the cluster generating algorithm and/or modifying the hierarchical structure and priors of the data model. Further research needs to be done to identify the root causes of this behavior. However, this problem is alleviated by the fact that, as shown in this study, type 1 errors of differential methylation tests can likely be adequately calibrated as long as the standard deviation of beta values (i.e. noise level) is identified. This information may even be obtained from the model fits themselves, since standard deviations of Z and X are already estimated via the model.

Cell specific results from simulation C suggest that statistical power for testing differences in mean methylation could be higher for the most abundant cell type compared to cell types with smaller cell proportions. With the exception of this finding, no clear trend of power as a function of cell proportion was observed after calibrating type 1 error. Furthermore, type 1 error rates for fixed credible interval decision rules appeared to vary but did not form a trend with cell proportion either. These two observations could potentially suggest that there are unaccounted properties of the cell types themselves that may affect rejection rules independently of their cell fractions. This could potentially mean that in order to achieve adequate calibration of type 1 errors further considerations and incorporating more information into the analysis is necessary. However, this issue was observed in both spatial and non-spatial models and is therefore not unique to our proposed modelling approach. Further research is needed investigating different means of cell proportions, different variances of cell proportions and different assignments of cell proportions to cell types in order to validate the relationships between blood cell type, cell proportion and testing performance.

Even though the prior distributions chosen in this study appear to represent a reasonable option, more work needs to be done in order to determine an optimal configuration, as alluded to earlier. The fact that performance in our simulations was highly sensitive to prior configurations supports this notion. So far informative restrictions were only imposed on variances of beta values, based on the knowledge that for response values bounded between 0 and 1 variances cannot exceed 0.25. However, following a similar thought process, informative priors could be chosen for mean methylation values and effect sizes that better acknowledge the restriction to the (0,1) interval. Different

priors for the shrinkage parameters $\xi^{[O]}, \xi^{[OO]}, \xi^{[OOO]}$ may also lead to further improvement. Even our selected informative type 2 priors are by no means guaranteed to be the most optimal choice and other prior shapes or prior families may be more suitable for variance parameters. The fact that honoring the restriction to the (0,1) interval appeared to improve performance does also suggest that employing a Beta distributed likelihood of the response instead of a Normal distributed likelihood may substantially improve performance and should be evaluated in future research.

In summary, our study shows that models leveraging spatial DNAm correlation between CpGs can improve model fit and statistical tests of cell-specific differential methylation. Our proposed hierarchical Bayesian approach to spatial modelling led to consistently lower prediction error, model complexity and higher statistical power after calibrating type 1 error rates than non-spatial variants of the same model. Our results suggest that the TCA approach by Rahmani et al. (2019), currently the most powerful approach for testing cell-specific differences in mean methylation based on heterogeneous bulk samples, could likely be improved by leveraging spatial correlation. Lastly, there is still a need to further refine our modelling strategy and to validate our findings in biological datasets where presence and absence of differential methylation are known.

Chapter 5

Summary and Future Directions

Analysis of compositional data, in which multiple individual components sum to a fixed sum, is complicated by negative correlation bias and emergence of spurious associations. Experiments that do not account for the presence of compositional effects are prone to identify non-reproducible, false-positive associations and may even miss true relationships between collected variables. Unfortunately, compositions naturally arise when relative information, such as proportions of a whole, are collected and compositional data is prevalent in various fields of research. Characteristics of biological samples are frequently constrained by compositionality, since they contain a multitude of complex structural and biochemical components that are limited in their number and in any signal they may emit by the total collected biomass. The development and application of statistical methods that acknowledge compositionality in the biomedical field has therefore remained an important objective in modern research.

Microbiome studies constitute a prominent example of this notion, since they inherently try to answer questions about how phenotypes associate with the overall microbial composition or with the abundance of individual microbes. In this dissertation, a variety of different methods were successfully applied to characterize the overall microbiome of pancreatic tissue from both pancreatic cancer and non-cancer subjects. Novel findings that samples in both groups were highly subject specific yet showed clear similarities between pancreatic and duodenum tissue help to inform future research aiming to understand connections and interactions between microbial communities of the gut. Curiously, pancreatic and duodenum tissues were also found to contain microbes commonly identified in the oral cavity. The successful identification of microbes for which abundance significantly differed either between cancer and non-cancer subjects and between the individual disease subtypes sheds light on how microbes may associate with disease. Key limitations of this study were the small number of tissue samples from cancer subjects, the small number of subjects for which multiple tissue samples were available, and the fact that samples of non-cancer subjects

were completely confounded with data source. Future studies collecting a larger number of samples originating from both healthy individuals, as well as subjects with different cancer subtypes, are necessary to validate our findings. If possible, the confounding issue encountered in our study should be resolved by balancing cancer and non-cancer subjects across multiple different clinical sites that perform analyses individually.

The initial study of the pancreatic microbiome also motivated the question of whether microbes of the oral cavity could potentially consistently associate with microbes of the gut with respect to some phenotype of interest. Such microbes may allow information about microbial composition in the gut to be inferred based on the evaluation of oral samples; an exciting prospect, considering samples of the gut can only be collected through invasive, surgical techniques. In the third chapter of this dissertation, we presented a novel Bayesian framework able to test for these types of associations while simultaneously encompassing a variety of different data models that are able to address the challenges associated with microbial abundance data. Simulation studies utilizing different data models and different degrees of association show that the approach is able to correctly identify patterns of interest and suggest that certain data models can achieve adequate power for a moderate degree of association and modest sample sizes. Since the approach involves estimating the posterior distribution of a given correlation statistic that was applied to posterior estimates of microbial abundance, future research should evaluate the impact that different choices of correlation statistics, in combination with different data models, could have on statistical inference. A limitation of the current approach is its potential lack of power when complex functions of population parameters are tested for association or low-level phylogeny data is considered. Future developments should aim to improve statistical power in these scenarios. Applying our method to samples from pancreatic cancer subjects led to the successful identification of microbes that exhibit consistent patterns with respect to cancer subtype between mouth and gut, two of which had previously been shown to be associated with pancreatic cancer in a different context. While these results are promising, more research is needed to validate our findings. Lastly, considering the generality of the proposed approach, we aim to potentially apply our testing framework to different

data models and different research settings in which similar types of associations are of interest.

In the fourth chapter of this work, we presented an extension of statistical approaches that estimate cell type specific CpG methylation based on bulk samples that contain compositions of multiple cell types. In contrast to previous approaches, our method leverages the well-known spatial correlation between CpGs by utilizing Bayesian hierarchical models. Specifically, after forming clusters of proximal CpGs, our method shrinks estimates of mean methylation towards an overall mean for each cluster, while also shrinking mean methylation estimates of proximal clusters towards each other. The overall approach, as well as tuning parameters employed during the cluster forming algorithm are closely informed by evaluating whole blood methylation data of isolated cell types. Extensive simulation studies show consistent improvements with regards to prediction accuracy and statistical power of our spatial models compared to non-spatial approaches, but also highlight potential issues with regards to inflation and calibration of type 1 errors. While the presented results show promise, our approach is almost certainly not fully optimized due to the complexity of both the biological data and the modelling strategy itself. Future developments should focus on assessing performance for different cluster forming strategies, such as approaches that are driven by annotation of known biological functions of CpGs, and for different hierarchical model structures. Additional simulations should also aim to extend to chromosomes other than 1 and 13, and to assess performance as a function of different sampling strategies of cell proportions in combination with their assignment to different cell types. Finally, validation by applying spatial and non-spatial models to other biological datasets for which both bulk methylation levels and cell type specific methylation levels are available, will be necessary. Such analyses could help to confirm whether the conducted simulations are representative and also whether the appproach could potentially be extended to settings involving different tissues with different cell types.

References

ACS (2018). Cancer facts and figures 2018. American Cancer Society, Inc., Atlanta.

Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., Riso, G. D., Monticelli, A., Miele, G., Chiariotti, L., & Cocozza, S. (2020). Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, 112(1), 144–150.

Al-Hebshi, N. N., Nasher, A. T., Idris, A. M., & Chen, T. (2015). Robust species taxonomy assignment algorithm for 16s rRNA NGS reads: application to oral carcinoma samples. *Journal of Oral Microbiology*, 7(1), 28934.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32–46.

Angelakis, E., Armougom, F., Carrière, F., Bachar, D., Laugier, R., Lagier, J.-C., Robert, C., Michelle, C., Henrissat, B., & Raoult, D. (2015). A metagenomic investigation of the duodenal microbiota reveals links with obesity. *PLOS ONE*, 10(9), e0137784.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics*, 30(10), 1363–1369.

Bergogne-Berezin, E. & Towner, K. J. (1996). Acinetobacter spp. as nosocomial pathogens: microbiological, clinical, and epidemiological features. *Clinical Microbiology Reviews*, 9(2), 148–165.

Berry, S. M. & Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60(2), 418–426.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), 127–135.

Bray, J. R. & Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4), 325–349.

Brook, I. & Frazier, E. H. (1996). Microbiological analysis of pancreatic abscess. *Clinical Infectious Diseases*, 22(2), 384–385.

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., Chipashvili, O., Hagan, T., Walker, M., Ramachandran, A., Diosdado, B., Serna, G., Mulet, N., Landolfi, S., y Cajal, S. R., Fasani, R., Aguirre, A. J., Ng, K., Élez, E., Ogino, S., Tabernero, J., Fuchs, C. S., Hahn, W. C., Nuciforo, P., & Meyerson, M. (2017). Analysis of fusobacterium persistence and antibiotic response in colorectal cancer. *Science*, 358(6369), 1443–1448.

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12), 2639–2643.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., & Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5), 335–336.

Castellarin, M., Warren, R. L., Freeman, J. D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R. A., & Holt, R. A. (2011). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Research*, 22(2), 299–306.

Catoni, M., Tsang, J. M., Greco, A. P., & Zabet, N. R. (2018). DMRcaller: a versatile r/bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Research*.

Chai, H., Jiang, H., Lin, L., & Liu, L. (2018). A marginalized two-part beta regression model for microbiome compositional data. *PLOS Computational Biology*, 14(7), e1006329.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12), 4185–4193.

Chen, C., Hemme, C., Beleno, J., Shi, Z. J., Ning, D., Qin, Y., Tu, Q., Jorgensen, M., He, Z., Wu, L., & Zhou, J. (2018). Oral microbiota of periodontal health and disease and their changes after nonsurgical periodontal therapy. *The ISME Journal*, 12(5), 1210–1224.

Chen, E. Z. & Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17), 2611–2617.

chiang Tsui, N., Zhao, E., Li, Z., Miao, B., Cui, Y., Shen, Y., & Qu, P. (2009). Microbiological findings in secondary infection of severe acute pancreatitis. *Pancreas*, 38(5), 499–502.

Chiranjeevi, T., Prasad, O. H., Prasad, U., Kumar, A. K., Chakravarthi, V., Rao, P. B., Sarma, P., Reddy, N., & Bhaskar, M. (2014). Identification of microbial pathogens in periodontal disease and diabetic patients of south indian population. *Bioinformation*, 10(4), 241–244.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R., Sugarbaker, D. J., Yeh, R.-F., Wiencke, J. K., & Kelsey, K. T. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLOS Genetics*, 5(8), e1000602.

Chung, M., Zhao, N., Meier, R., Koestler, D. C., Wu, G., Castillo, E. D., Paster, B. J., Kelsey, K. T., & Michaud, D. S. (2019). Oral, gut, and pancreatic microbiome are correlated and exhibit consist co-abundance in patients with pancreatic diseases and cancer. *[manuscript in progress]*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

del Castillo, E., Meier, R., Chung, M., Koestler, D. C., Chen, T., Paster, B. J., Charpentier, K. P., Kelsey, K. T., Izard, J., & Michaud, D. S. (2019). The microbiomes of pancreatic and duodenum tissue overlap and are highly subject specific but differ between pancreatic cancer and noncancer subjects. *Cancer Epidemiology Biomarkers & Prevention*, 28(2), 370–383.

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., Lakshmanan, A., & Wade, W. G. (2010). The human oral microbiome. *Journal of Bacteriology*, 192(19), 5002–5017.

Downes, J., Vartoukian, S. R., Dewhirst, F. E., Izard, J., Chen, T., Yu, W.-H., Sutcliffe, I. C., & Wade, W. G. (2009). Pyramidobacter piscolens gen. nov., sp. nov., a member of the phylum synergistetes isolated from the human oral cavity. *International Journal of Systematic and Evolutionary Microbiology*, 59(5), 972–980.

Elliott, D. R. F., Walker, A. W., ODonovan, M., Parkhill, J., & Fitzgerald, R. C. (2017). A non-endoscopic device to sample the oesophageal microbiota: a case-control study. *The Lancet Gastroenterology & Hepatology*, 2(1), 32–42.

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., Clemente, J. C., Knight, R., Heath, A. C., Leibel, R. L., Rosenbaum, M., & Gordon, J. I. (2013). The long-term stability of the human gut microbiota. *Science*, 341(6141), 1237439.

Fan, X., Alekseyenko, A. V., Wu, J., Peters, B. A., Jacobs, E. J., Gapstur, S. M., Purdue, M. P., Abnet, C. C., Stolzenberg-Solomon, R., Miller, G., Ravel, J., Hayes, R. B., & Ahn, J. (2016). Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut*, 67(1), 120–127.

Fangous, M.-S., Hémon, F., Graf, P., Samier-Guérin, A., Alavi, Z., Bars, H. L., & Berre, R. L. (2016). Bone infections caused by gemella haemolysans. *Médecine et Maladies Infectieuses*, 46(8), 449–452.

Fardini, Y., Wang, X., Témoin, S., Nithianantham, S., Lee, D., Shoham, M., & Han, Y. W. (2011). Fusobacterium nucleatum adhesin FadA binds vascular endothelial cadherin and alters endothelial integrity. *Molecular Microbiology*, 82(6), 1468–1480.

Faveri, M., Figueiredo, L. C., Duarte, P. M., Mestnik, M. J., Mayer, M. P. A., & Feres, M. (2009). Microbiological profile of untreated subjects with localized aggressive periodontitis. *Journal of Clinical Periodontology*, 36(9), 739–749.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.

Fortin, J.-P., Triche, T. J., & Hansen, K. D. (2016). Preprocessing, normalization and integration of the illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, (pp. btw691).

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., Giannoukos, G., Boylan, M. R., Ciulla, D., Gevers, D., Izard, J., Garrett, W. S., Chan, A. T., & Huttenhower, C. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences*, 111(22), E2329–E2338.

García-Lechuz, J. M., Cuevas-Lobato, O., Hernángomez, S., Hermida, A., Guinea, J., Marín, M., Peláez, T., & Bouza, E. (2002). Extra-abdominal infections attributable to gemella species. *International Journal of Infectious Diseases*, 6(1), 78–82.

Geller, L. T., Barzily-Rokni, M., Danino, T., Jonas, O. H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z. A., Shee, K., Thaiss, C. A., Reuben, A., Livny, J., Avraham, R., Frederick, D. T., Ligorio, M., Chatman, K., Johnston, S. E., Mosher, C. M., Brandis, A., Fuks, G., Gurbatri, C., Gopalakrishnan, V., Kim, M., Hurd, M. W., Katz, M., Fleming, J., Maitra, A., Smith, D. A., Skalak, M., Bu, J., Michaud, M., Trauger, S. A., Barshack, I., Golan, T., Sandbank, J., Flaherty, K. T., Mandinova, A., Garrett, W. S., Thayer, S. P., Ferrone, C. R., Huttenhower, C., Bhatia, S. N., Gevers, D., Wargo, J. A., Golub, T. R., & Straussman, R. (2017). Potential role of

intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*, 357(6356), 1156–1160.

Gibson, F., Yumoto, H., Takahashi, Y., Chou, H.-H., & Genco, C. (2006). Innate immune signaling and porphyromonas gingivalis-accelerated atherosclerosis. *Journal of Dental Research*, 85(2), 106–121.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8.

Gonçalves, L. F. H., Fermiano, D., Feres, M., Figueiredo, L. C., Teles, F. R. P., Mayer, M. P. A., & Faveri, M. (2012). Levels of selenomonas species in generalized aggressive periodontitis. *Journal of Periodontal Research*, 47(6), 711–718.

Goodman, B. & Gardner, H. (2018). The microbiome and cancer. *The Journal of Pathology*, 244(5), 667–676.

Gur, C., Ibrahim, Y., Isaacson, B., Yamin, R., Abed, J., Gamliel, M., Enk, J., Bar-On, Y., Stanietsky-Kaynan, N., Coppenhagen-Glazer, S., Shussman, N., Almogy, G., Cuapio, A., Hofer, E., Mevorach, D., Tabib, A., Ortenberg, R., Markel, G., Miklić, K., Jonjic, S., Brennan, C. A., Garrett, W. S., Bachrach, G., & Mandelboim, O. (2015). Binding of the fap2 protein of fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity*, 42(2), 344–355.

Hajishengallis, G. (2014). Immunomicrobial pathogenesis of periodontitis: keystones, pathobionts, and host response. *Trends in Immunology*, 35(1), 3–11.

Hawinkel, S., Kerckhof, F.-M., Bijnens, L., & Thas, O. (2019). A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLOS ONE*, 14(2), e0205474.

Hieken, T. J., Chen, J., Hoskin, T. L., Walther-Antonio, M., Johnson, S., Ramaker, S., Xiao, J., Radisky, D. C., Knutson, K. L., Kalari, K. R., Yao, J. Z., Baddour, L. M., Chia, N., & Degnim, A. C. (2016). The microbiome of aseptically collected human breast tissue in benign and malignant disease. *Scientific Reports*, 6(1).

Hill, M., Dach, J., Barkin, J., Isikoff, M., & Morse, B. (1983). The role of percutaneous aspiration in the diagnosis of pancreatic abscess. *American Journal of Roentgenology*, 141(5), 1035–1038.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., & Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1).

Ishiya, K. & Aburatani, S. (2019). Outlier detection for minor compositional variations in taxonomic abundance data. *Applied Sciences*, 9(7), 1355.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, 41(1), 200–209.

Jin, Z. & Liu, Y. (2018). DNA methylation in human diseases. *Genes & Diseases*, 5(1), 1–8.

Klein, A. P., Lindström, S., Mendelsohn, J. B., Steplowski, E., Arslan, A. A., de Mesquita, H. B. B., Fuchs, C. S., Gallinger, S., Gross, M., Helzlsouer, K., Holly, E. A., Jacobs, E. J., LaCroix, A., Li, D., Mandelson, M. T., Olson, S. H., Petersen, G. M., Risch, H. A., Stolzenberg-Solomon, R. Z., Zheng, W., Amundadottir, L., Albanes, D., Allen, N. E., Bamlet, W. R., Boutron-Ruault, M.-C., Buring, J. E., Bracci, P. M., Canzian, F., Clipp, S., Cotterchio, M., Duell, E. J., Elena, J., Gaziano, J. M., Giovannucci, E. L., Goggins, M., Hallmans, G., Hassan, M., Hutchinson, A., Hunter, D. J., Kooperberg, C., Kurtz, R. C., Liu, S., Overvad, K., Palli, D., Patel, A. V., Rabe, K. G., Shu, X.-O., Slimani, N., Tobias, G. S., Trichopoulos, D., Eeden, S. K. V. D., Vineis, P., Virtamo, J., Wactawski-Wende, J., Wolpin, B. M., Yu, H., Yu, K., Zeleniuch-Jacquotte, A., Chanock, S. J., Hoover, R. N., Hartge, P., & Kraft, P. (2013). An absolute risk model to identify

individuals at elevated risk for pancreatic cancer in the general population. *PLOS ONE*, 8(9), e72311.

Koestler, D. C., Jones, M. J., Usset, J., Christensen, B. C., Butler, R. A., Kobor, M. S., Wiencke, J. K., & Kelsey, K. T. (2016). Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics*, 17(1).

Koestler, D. C., Marsit, C. J., Christensen, B. C., Accomando, W., Langevin, S. M., Houseman, E. A., Nelson, H. H., Karagas, M. R., Wiencke, J. K., & Kelsey, K. T. (2012). Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. *Cancer Epidemiology Biomarkers & Prevention*, 21(8), 1293–1302.

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., Clancy, T. E., Chung, D. C., Lochhead, P., Hold, G. L., El-Omar, E. M., Brenner, D., Fuchs, C. S., Meyerson, M., & Garrett, W. S. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe*, 14(2), 207–215.

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., Baselga, J., Liu, C., Shivdasani, R. A., Ogino, S., Birren, B. W., Huttenhower, C., Garrett, W. S., & Meyerson, M. (2011). Genomic analysis identifies association of fusobacterium with colorectal carcinoma. *Genome Research*, 22(2), 292–298.

LaMonte, M. J., Genco, R. J., Hovey, K. M., Wallace, R. B., Freudenheim, J. L., Michaud, D. S., Mai, X., Tinker, L. F., Salazar, C. R., Andrews, C. A., Li, W., Eaton, C. B., Martin, L. W., & Wactawski-Wende, J. (2017). History of periodontitis diagnosis and edentulism as predictors of cardiovascular disease, stroke, and mortality in postmenopausal women. *Journal of the American Heart Association*, 6(4).

Li, S., Fuhler, G. M., BN, N., Jose, T., Bruno, M. J., Peppelenbosch, M. P., & Konstantinov, S. R. (2017). Pancreatic cyst fluid harbors a unique microbiome. *Microbiome*, 5(1).

Liu, B., Faller, L. L., Klitgord, N., Mazumdar, V., Ghodsi, M., Sommer, D. D., Gibbons, T. R., Treangen, T. J., Chang, Y.-C., Li, S., Stine, O. C., Hasturk, H., Kasif, S., Segrè, D., Pop, M., & Amar, S. (2012). Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PLOS ONE*, 7(6), e37919.

Liu, Y., Li, X., Aryee, M. J., Ekström, T. J., Padyukov, L., Klareskog, L., Vandiver, A., Moore, A. Z., Tanaka, T., Ferrucci, L., Fallin, M. D., & Feinberg, A. P. (2014). GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *The American Journal of Human Genetics*, 94(4), 485–495.

Lu, H., Ren, Z., Li, A., Li, J., Xu, S., Zhang, H., Jiang, J., Yang, J., Luo, Q., Zhou, K., Zheng, S., & Li, L. (2019). Tongue coating microbiome data distinguish patients with pancreatic head cancer from healthy controls. *Journal of Oral Microbiology*, 11(1), 1563409.

Martino, C., Morton, J. T., Marotz, C. A., Thompson, L. R., Tripathi, A., Knight, R., & Zengler, K. (2019). A novel sparse compositional technique reveals microbial perturbations. *mSystems*, 4(1).

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2011). An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3), 610–618.

Meier, R. (2019). R scripts for simulation studies, analyses and examples related to pasta. `https://github.com/richard-meier/PASTA_scripts`. [Online; accessed 08-April-2019].

Michaud, D. S. (2013). Role of bacterial infections in pancreatic cancer. *Carcinogenesis*, 34(10), 2193–2197.

Michaud, D. S., Fu, Z., Shi, J., & Chung, M. (2017). Periodontal disease, tooth loss, and cancer risk. *Epidemiologic Reviews*, 39(1), 49–58.

Michaud, D. S., Izard, J., Wilhelm-Benartzi, C. S., You, D.-H., Grote, V. A., Tjønneland, A., Dahm, C. C., Overvad, K., Jenab, M., Fedirko, V., Boutron-Ruault, M. C., Clavel-Chapelon, F., Racine, A., Kaaks, R., Boeing, H., Foerster, J., Trichopoulou, A., Lagiou, P., Trichopoulos, D., Sacerdote, C., Sieri, S., Palli, D., Tumino, R., Panico, S., Siersema, P. D., Peeters, P. H., Lund, E., Barricarte, A., Huerta, J.-M., Molina-Montes, E., Dorronsoro, M., Quirós, J. R., Duell, E. J., Ye, W., Sund, M., Lindkvist, B., Johansen, D., Khaw, K.-T., Wareham, N., Travis, R. C., Vineis, P., de Mesquita, H. B. B., & Riboli, E. (2012a). Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large european prospective cohort study. *Gut*, 62(12), 1764–1770.

Michaud, D. S., Izard, J., Wilhelm-Benartzi, C. S., You, D.-H., Grote, V. A., Tjønneland, A., Dahm, C. C., Overvad, K., Jenab, M., Fedirko, V., Boutron-Ruault, M. C., Clavel-Chapelon, F., Racine, A., Kaaks, R., Boeing, H., Foerster, J., Trichopoulou, A., Lagiou, P., Trichopoulos, D., Sacerdote, C., Sieri, S., Palli, D., Tumino, R., Panico, S., Siersema, P. D., Peeters, P. H., Lund, E., Barricarte, A., Huerta, J.-M., Molina-Montes, E., Dorronsoro, M., Quirós, J. R., Duell, E. J., Ye, W., Sund, M., Lindkvist, B., Johansen, D., Khaw, K.-T., Wareham, N., Travis, R. C., Vineis, P., de Mesquita, H. B. B., & Riboli, E. (2012b). Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large european prospective cohort study. *Gut*, 62(12), 1764–1770.

Mitsuhashi, K., Nosho, K., Sukawa, Y., Matsunaga, Y., Ito, M., Kurihara, H., Kanno, S., Igarashi, H., Naito, T., Adachi, Y., Tachibana, M., Tanuma, T., Maguchi, H., Shinohara, T., Hasegawa, T., Imamura, M., Kimura, Y., Hirata, K., Maruyama, R., Suzuki, H., Imai, K., Yamamoto, H., & Shinomura, Y. (2015). Association of fusobacterium species in pancreatic cancer tissues with molecular features and prognosis. *Oncotarget*, 6(9).

Moarii, M., Boeva, V., Vert, J.-P., & Reyal, F. (2015). Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, 16(1).

Mosquera, J., Zabalza, M., Laniero, M., & Blanco, J. (2000). Endocarditis due to gemella haemolysans in a patient with hemochromatosis. *Clinical Microbiology and Infection*, 6(10), 566–568.

Nilsson, H.-O. (2006). Helicobacter species ribosomal DNA in the pancreas, stomach and duodenum of pancreatic cancer patients. *World Journal of Gastroenterology*, 12(19), 3038.

Pan, F., Zhang, L., Li, M., Hu, Y., Zeng, B., Yuan, H., Zhao, L., & Zhang, C. (2018). Predominant gut lactobacillus murinus strain mediates anti-inflammaging effects in calorie-restricted mice. *Microbiome*, 6(1).

Pawlowsky-Glahn, V. & Egozcue, J. J. (2016). Spatial analysis of compositional data: A historical review. *Journal of Geochemical Exploration*, 164, 28–32.

Peng, X., Li, G., & Liu, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *Journal of Computational Biology*, 23(2), 102–110.

Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C. P., Kurz, E., Mishra, A., Mohan, N., Aykut, B., Usyk, M., Torres, L. E., Werba, G., Zhang, K., Guo, Y., Li, Q., Akkad, N., Lall, S., Wadowski, B., Gutierrez, J., Rossi, J. A. K., Herzog, J. W., Diskin, B., Torres-Hernandez, A., Leinwand, J., Wang, W., Taunk, P. S., Savadkar, S., Janal, M., Saxena, A., Li, X., Cohen, D., Sartor, R. B., Saxena, D., & Miller, G. (2018). The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discovery*, 8(4), 403–416.

Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., & Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9).

Quinn, T. P., Erb, I., Richardson, M. F., & Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16), 2870–2878.

Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., Rosset, S., Sankararaman, S., & Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature Communications*, 10(1).

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D., Söderhäll, C.,

Scheynius, A., & Kere, J. (2012). Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PLOS ONE*, 7(7), e41361.

Rogers, M. B., Aveson, V., Firek, B., Yeh, A., Brooks, B., Brower-Sinning, R., Steve, J., Banfield, J. F., Zureikat, A., Hogg, M., Boone, B. A., Zeh, H. J., & Morowitz, M. J. (2017). Disturbances of the perioperative microbiome across multiple body sites in patients undergoing pancreaticoduodenectomy. *Pancreas*, 46(2), 260–267.

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating e-cadherin/-catenin signaling via its FadA adhesin. *Cell Host & Microbe*, 14(2), 195–206.

Salas, L. A., Koestler, D. C., Butler, R. A., Hansen, H. M., Wiencke, J. K., Kelsey, K. T., & Christensen, B. C. (2018). An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the illumina HumanMethylationEPIC BeadArray. *Genome Biology*, 19(1).

Sarria, J. C., Vidal, A. M., & III, R. C. K. (2001). Infections caused by kluyvera species in humans. *Clinical Infectious Diseases*, 33(7), e69–e74.

Scheithauer, B. K., Wos-Oxley, M. L., Ferslev, B., Jablonowski, H., & Pieper, D. H. (2009). Characterization of the complex bacterial communities colonizing biliary stents reveals a host-dependent diversity. *The ISME Journal*, 3(7), 797–807.

Schmid, S. W., Uhl, W., Friess, H., Malfertheiner, P., & Buchler, M. W. (1999). The role of infection in acute pancreatitis. *Gut*, 45(2), 311–311.

Schneider, J., Schenk, P., Obermeier, A., Fremd, J., Feihl, S., Forkl, S., Wantia, N., Römmler, F., Neu, B., Bajbouj, M., von Delius, S., Schmid, R. M., Algül, H., & Weber, A. (2015). Microbial colonization of pancreatic duct stents. *Pancreas*, 44(5), 786–790.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients. *Anesthesia & Analgesia*, 126(5), 1763–1768.

Scola, B. L. & Raoult, D. (1998). Molecular identification of gemella species from three patients with endocarditis. *Journal of Clinical Microbiology*, 36(4), 866–871.

Segata, N., Haake, S., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., Huttenhower, C., & Izard, J. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology*, 13(6), R42.

Shi, P., Zhang, A., & Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2), 1019–1040.

Stingu, C.-S., Schaumann, R., Jentsch, H., Eschrich, K., Brosteanu, O., & Rodloff, A. C. (2012). Association of periodontitis with increased colonization byPrevotella nigrescens. *Journal of Investigative and Clinical Dentistry*, 4(1), 20–25.

Swidsinski, A. (2005). Bacterial biofilm within diseased pancreatic and biliary tracts. *Gut*, 54(3), 388–395.

Teschendorff, A. E. & Zheng, S. C. (2017). Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5), 757–768.

Titus, A. J., Gallimore, R. M., Salas, L. A., & Christensen, B. C. (2017). Cell-type deconvolution from DNA methylation: a review of recent applications. *Human Molecular Genetics*, 26(R2), R216–R224.

Tsilimigras, M. C. & Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, 26(5), 330–335.

Wang, L., Zhou, J., Xin, Y., Geng, C., Tian, Z., Yu, X., & Dong, Q. (2016). Bacterial overgrowth and diversification of microbiota in gastric cancer. *European Journal of Gastroenterology & Hepatology*, 28(3), 261–266.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wu, T., Zhang, Z., Liu, B., Hou, D., Liang, Y., Zhang, J., & Shi, P. (2013). Gut microbiota dysbiosis and bacterial community assembly associated with cholesterol gallstones in large-scale study. *BMC Genomics*, 14(1), 669.

Xia, Y., Sun, J., & Chen, D.-G. (2018). Modeling zero-inflated microbiome data. In *Statistical Analysis of Microbiome Data with R* (pp. 453–496). Springer Singapore.

Xu, Z., Niu, L., Li, L., & Taylor, J. A. (2015). ENmix: a novel background correction method for illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*, 44(3), e20–e20.

Ye, F., Shen, H., Li, Z., Meng, F., Li, L., Yang, J., Chen, Y., Bo, X., Zhang, X., & Ni, M. (2016). Influence of the biliary system on biliary bacteria revealed by bacterial communities of the human biliary and upper digestive tracts. *PLOS ONE*, 11(3), e0150519.

Yu, G., Gail, M. H., Consonni, D., Carugno, M., Humphrys, M., Pesatori, A. C., Caporaso, N. E., Goedert, J. J., Ravel, J., & Landi, M. T. (2016). Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features. *Genome Biology*, 17(1).

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1).

Zheng, S. C., Breeze, C. E., Beck, S., & Teschendorff, A. E. (2018). Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods*, 15(12), 1059–1066.

Appendix A

Declarations and Supplementary Material for Chapter 1

Declarations

Disclosure of Potential Conflicts of Interest

K.T. Kelsey is a consultant/advisory board member at Celintec. No potential conflicts of interest were disclosed by the other authors.

Availability of data and material

The datasets analyzed for this manuscript will be available through the NCBI under the BioProject accession number: PRJNA421501. R codes generated to analyze these data will be available upon request.

Acknowledgments

The research reported in this publication was supported by the NIH/NCI grants R01 CA166150 and P30 CA168524.

We thank the participants who graciously enrolled in this study. We thank Ms. Priyanka Joshi for her tremendous help with recruitment of subjects at the RIH, and Dr. Ross Taliano for assisting with the preparation of the tissue specimens in the Pathology Department at the RIH. We also thank Drs. Murray Resnick, Kara Lombardo, Alexis Kokaras, Emily Walsh, and Ms. Laura Gantt and Naisi Zhao for their help with this project.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Supplementary Methods

DNA extraction for tissue samples

All the tissue samples were homogenized using a Fast Prep instrument with lysing matrix Y (MP Biomedicals, CA, USA) after addition of 180μl of ATL buffer (QIAGEN, CA, USA). Bead-beating was performed for one minute at 6.0 m sec-1. The supernatant was then recovered and 200μl of lysozyme (20mg/ml) buffer was added. The samples were then incubated at 37°C for 30 min. Subsequently, 20-μl of proteinase K (20 mg/ml) was added and samples were further incubated at 56°C for 30 min. Afterwards, 200-μl lysis buffer AL (QIAGEN) was added and samples were incubated at 56°C for 30 min. Subsequently, 900-μl of 4M guanidine thiocyanate buffer was added, samples were mixed by inversion. Subsequently, 700-μl cold ethanol was added and the DNA was purified using the DNeasy Blood and Tissue kit (69506, QIAGEN) as per the manufacturer's tissue protocol (DNeasy Blood & Tissue Handbook, version 07/2006). DNA concentration was fluorometrically measured using the QuantiT PicoGreen dsDNA High-Sensitivity Assay (Q-33120, Life Technologies) with a BioTek fluorescence plate reader instrument (Ex.λ/Em.λ of 485/548 nm) using the BioTek Gen5 software package.

Bacterial amplification

Bacterial amplification of all the DNA extracted from samples was validated using primers targeting the V3-V4 hypervariable region of the 16S ribosomal RNA gene with identical sequences designed for Illumina sequencing (Primers: 341F 5'-CCTACGGGAGGCAGCAG-3' and 806R 5'GGACTACHVGGGTWTCTAAT-3') without the addition of Illumina adapters and sample barcode sequencing.

Bacterial mock community

Genomic DNA from ten bacterial strains were used to build a bacterial mock community that was sequenced as a positive control in every MiSeq run performed throughout this research. The ten

bacterial strains that comprised this bacterial mock community were as follows: Cryptobacterium curtum Oral Taxon 579, Bacteroidales [G-2] sp. Oral Taxon 274, Capnocytophaga sp. Oral Taxon 338, Streptococcus anginosus Oral Taxon 543, Peptoniphilus sp. Oral Taxon 386, Selenomonas noxia Oral Taxon 130, Fusobacterium nucleatum ss polymorphum Oral Taxon 202, Aggregatibacter aphrophilus Oral Taxon 545, and Pyramidobacter piscolens Oral Taxon 357. Results for relative abundance of these control samples were consistent, and as expected, across each MiSeq run.

PCR

PCR mixtures of 50-μl contained 10-μl of diluted DNA template, 20-μl of HotMasterMix, 1-μl of each primer (10μM). The cycling conditions consisted of an initial template denaturation of 94°C for 3 min, followed by 30-cycles of denaturation at 94°C for 45 sec, annealing at 50°C for 60 sec, extension at 72°C for 1.5 min, and a final extension at 72°C for 10 min. Five-microliters of each PCR product loaded with gel loading dye were run on a 1% agarose gel in 1X Tris-acetate-EDTA (TAE) buffer stained with Sybr Safe DNA and visualized using an Alpha-Innotech instrument equipped with the FluorChem Q imaging software (Ex. λ/Em. λ of 475/537nm).

16S rRNA gene Amplicon Illumina Sequencing

10-50 ng of each metagenomic DNA template was first amplified using the sequencing primers designed to incorporate Illumina adapters and a sample barcode sequence, allowing directional sequencing covering the hypervariable region V3-V4. Primers used were as follows: 341F (AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGTCCTACGGGAGGCAGCAG) and 806R (CAAGCAGAAGACGGCATACGAGATN NNNNNNNNNNNAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT) (sequences of the primers are in italics, and N sequences corresponding to the barcodes). PCR mixtures contained 10-μl of diluted DNA template, 10-μl of HotMaster Taq DNA Polymerase Mix (5 Prime), and 1μl of each primer mix (10 μM). The cycling conditions consisted of an initial denaturation of 94°C for 3 min, followed by 30 cycles of denaturation at 94°C for 45 sec, annealing at 50 °C for 60 sec, extension at 72°C for 1.5 min, and a final extension

at 72°C for 10 min.

PCR products were then purified using a magnetic bead capture kit Agencourt Ampure XP purification beads (Beckman Coulter, Brea, CA, USA). Amplicons from each library were quantified and pooled in equimolar concentrations. Pooled libraries were electrophoresed in a 2% agarose gel with gel loading dye and Sybr Safe DNA gel stain. Bands were visualized under under UV transillumination, the band at 590 bp was excised and DNA was purified using the Minelute Gel Extraction kit (Qiagen). The purified DNA libraries pool was quantitated on an Agilent bioanalyzer DNA 1000 chips (Agilent, Santa Clara, CA, USA) using a Bioanalyzer to verify the DNA size fragment. The final concentration of the library was determined using a SYBR green quantitative PCR (qPCR) assay with primers specific to the Illumina adapters (Kapa Biosystems, Woburn, MA, USA) using a LightCycler 96 Real-Time PCR System Roche Diagnostics GmbH, Mannheim, Germany). The final amplicon pool was denatured at 4 nM before diluting to a final concentration of 12 pM. The libraries pool was then mixed with >5% PhiX Illumina control and were sequenced by 2 x 250 bp paired-end sequencing on the Miseq platform using MiSeq V2 reagent kit (Illumina, CA, USA), according to the manufacturer's specifications and generating paired-end reads of 250b in length in each direction.

Statistical analysis

In addition, PERMANOVA tests were conducted to compare beta-diversity measures (i.e., Bray-Curtis) between sites (i.e., pancreatic duct, pancreatic tail, pancreatic head, etc.), and sample groups (i.e., disease versus non-diseased, etc.). Briefly, PERMANOVA is an extension of the traditional analysis of variance (ANOVA) to a square matrix of pairwise distances with significance testing performed by permutation (Anderson, 2001).

Zero-inflated beta regression models represent a general class of mixture models where the response variable is assumed to have mixed continuous-discrete distribution with probability mass at zero. For our application, a logistic regression component to model OTU presence/absence (p0) and a beta regression component was used to model non-zero microbial abundance (μ). The

rationale for selecting this model stems from two distinct characteristics of microbiome data: the preponderance of zero OTU counts across samples (Chen & Li, 2016), commonly referred to as zero-inflation, and the fact that OTU relative abundance measurements are continuous and bounded between 0 and 1, and as a result, are reasonably well approximated with a beta distribution. Zero-inflated beta regression models were fit using the function "BEINF0", as implemented in the R package "gamlss".

Due to sample size limitations, associations between genus-level relative abundances and demographic/clinical variables were identified marginally by fitting zero-inflated beta regression models regressing on a single predictor. Models were fit only to genera with less than 90% of the counts being 0. As such, the total number of genera that were tested thus varied across the considered model. Models failing to converge due to data sparseness were considered not significant and were not carried forward for subsequent analyses. Associations were identified by conducting likelihood-ratio tests (LRT) and considered potentially meaningful when either the LRT p-value was less than 0.05 or the Akaike Information Criterion (AIC) of the alternative model was smaller when compared to the null model.

We conducted statistical analyses focused on identifying genera for which relative abundance differed significantly between RIH cancer patients and NDRI non-cancer patients across the set of pancreatic sites. Models were fit to the OTU data from the following tissue samples: pancreatic duct, pancreatic head, pancreatic tail, pancreatic tumor, pancreatic normal and duodenum. In order to account for within-subject correlation, a random intercept term for subject IDs was incorporated into the zero-inflated beta regression models. The utilized models also adjusted for age, sex, and log library size as fixed effects. Other multi-level categorical predictors such as sequencing run or body site were not included into the models in order to reduce sparseness and improve convergence behavior. Unfortunately, testing fixed effects in generalized linear mixed models via simple LRTs is known to be inefficient and unreliable for small to moderate sample sizes (Bolker et al., 2009). To address this issue, permutation based tests were utilized. First, the observed likelihood ratio statistic (LRS) comparing the full against the null model, excluding study ID as covariate, was

calculated. The null distribution of the LRS was then estimated by permuting study ID labels across patients. P-values were derived from 500 permutations per genus and adjusting for multiple testing was achieved via the false discovery rate method.

We also considered zero-inflated beta regression to compare relative abundance of bacteria by ICD code to evaluate whether profiles differed across the different types of RIH patients. For the purpose of this analysis, ICD10 codes were grouped into three categories: pancreatic cancer (ICD10 codes C25.0-C25.9), periampullary cancer (ICD10 codes C24.0-C24.1), and other pancreatic conditions (ICD10 codes K86.0-K86.3) (Table 2.1). We hereafter refer to these categories as C25, C24, and K86, respectively. Two strategies were considered: First, 30 NDRI pancreatic-head samples were compared with 30 RIH tumor-samples and adjusted for age, BMI, sex and sequencing run. The same strategy was also applied to compare the effects of ICD codes in RIH samples from NDRI samples using data from duodenum and pancreatic duct tissue. In the second approach, we restricted the analysis to RIH patients to account for other clinical variables (e.g. prior chemotherapy or use of antibiotics in prior 6 months); covariates that were adjusted for were selected empirically by identifying variables that exhibited an association with at least 30% of the OTUs that were formally tested. Bonferroni corrections were made to adjust for multiple comparisons when interpreting relative mean abundances of genera across the ICD codes; p-values from the Wald-tests for the mean value comparisons for the ICD codes were considered meaningful if they were less than 0.00057 (0.05/88; given that a maximum of 88 genera were tested).

Supplemental Figures

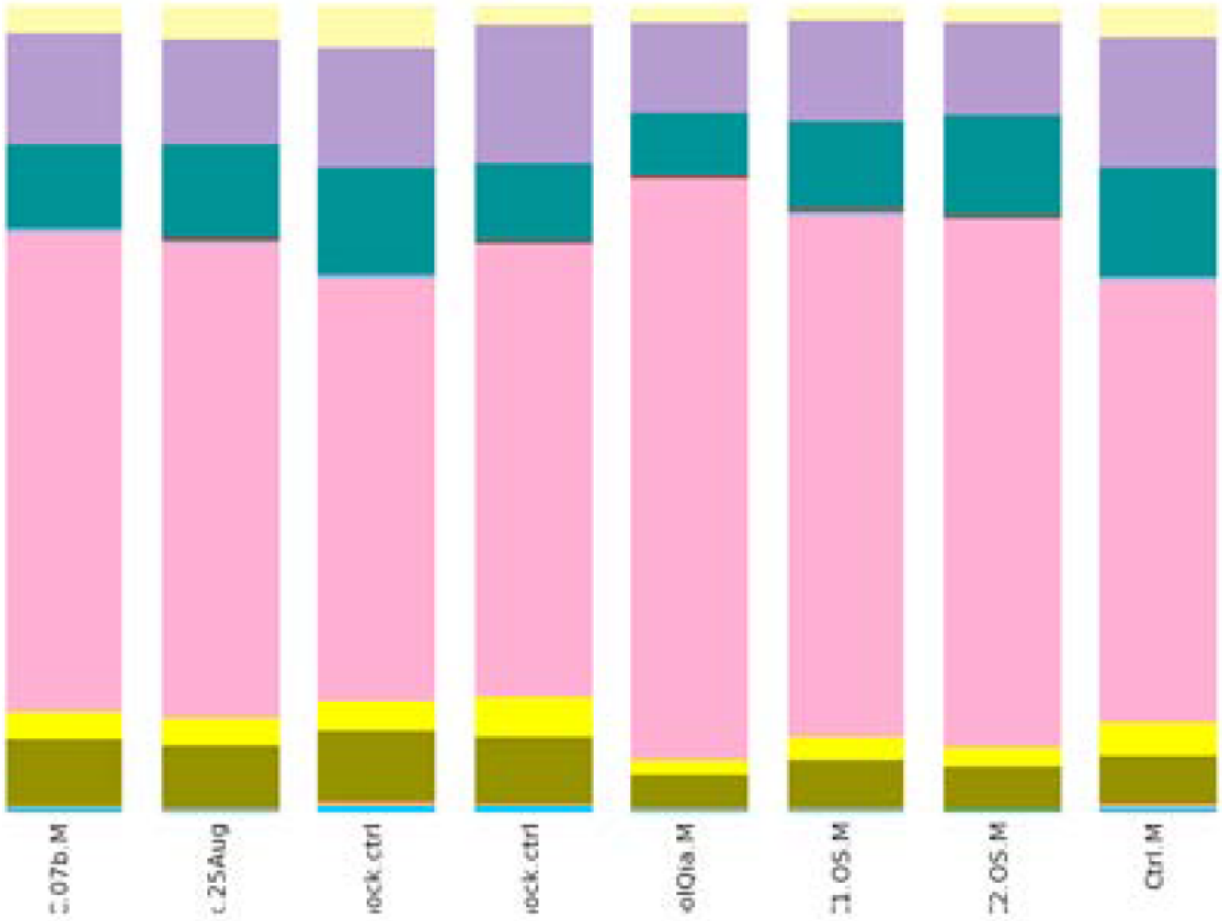
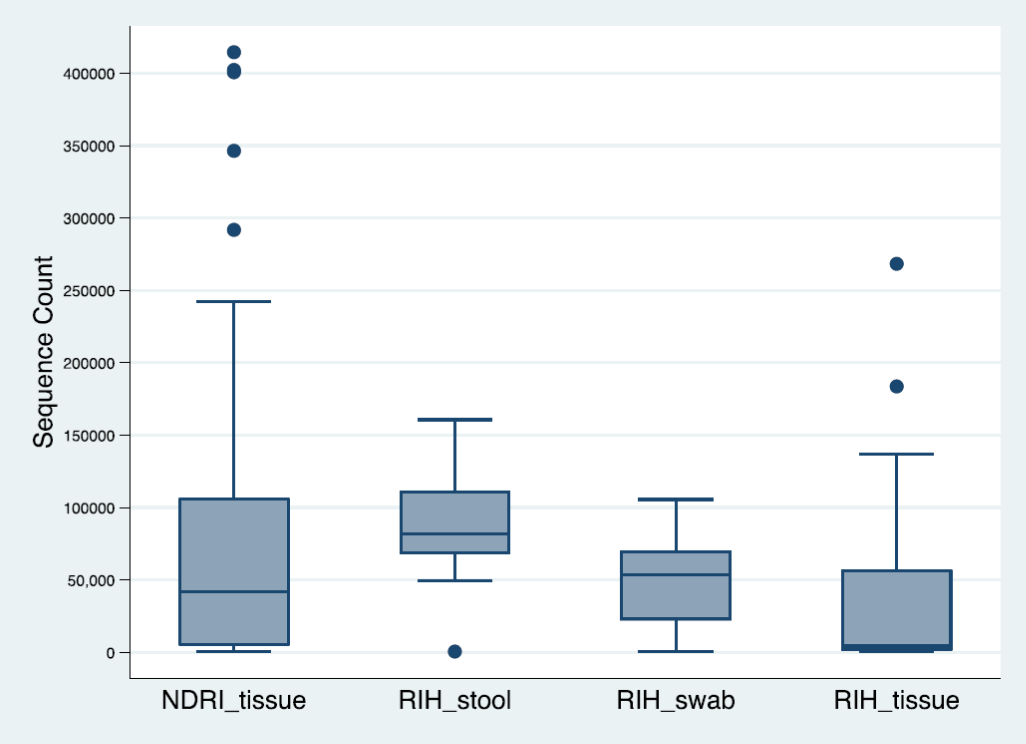


Figure A.1: Bacterial taxonomy (genera level) for the control samples of bacterial mock communities included in each of the MiSeq runs for this project.



Sample source (n)	Sequence count: median (Interquartile range)
NDRI tissue samples (113)	41,930 (4781 - 106,585)
RIH swab samples (57)	53,670 (22,231 - 70,376)
RIH tissue samples (76)	4342 (1025 – 57,130)
RIH stool samples (12)	82,018 (67,934 - 111,447

Figure A.2: Range of sequence counts in all the samples (after rarefaction at 500 counts).

Supplemental Tables

Table A.1: Number of samples per anatomical site from the Rhode Island Hospital [RIH] and the National Disease Research Interchange [NDRI].

Source	Anatomical Site	N
RIH	Bile Duct Swab	20
RIH	Duodenum	17
RIH	Jejunum Swab	31
RIH	Normal Pancreas	6
RIH	Pancreas Tumor	31
RIH	Pancreatic Duct	22
RIH	Stomach Swab	6
RIH	Stool	12
NDRI	Duodenum	32
NDRI	Pancreas Head	30
NDRI	Pancreas Tail	19
NDRI	Pancreatic Duct	32
Total		258

Table A.2: Results (at the species level) from multivariable zero-inflated beta regression models comparing bacteria presence/absence and relative abundance in tissue and swab samples from NDRI and RIH subjects*

Species	Total Read Counts	Non-zero Samples	Estimated Mean Relative Abundance (μ)**			Estimated Proportion of Presence (P1)			Global Perm Test*		
			RIH	NDRI	Wald p-value	RIH	NDRI	Wald p-value	p-value	p-adjusted^	AIC difference
G.multispecies_spp670_3	6160	27	0.0004	< 0.0001	< 0.0001	0.1238	0.0050	0.0001	< 0.0001	< 0.0001	5.32
L. gasseri	526388	120	0.0090	0.0144	0.0328	0.2598	0.8260	< 0.0001	< 0.0001	< 0.0001	21.78
L. salivarius	354859	111	0.0068	0.0202	0.0000	0.2767	0.6752	< 0.0001	< 0.0001	< 0.0001	13.92
S. intermedius	1295	52	0.0003	0.0003	0.2429	0.4586	0.1190	0.0003	< 0.0001	< 0.0001	1.55
S. multispecies_spp573_2	16829	44	0.0021	0.0007	0.0170	0.1662	0.0110	0.0000	< 0.0001	< 0.0001	7.52
L. saburreum	327	25	0.0004	0.0012	0.0026	0.0725	0.0019	0.0001	< 0.0001	< 0.0001	6.24
A. vaginalis	47175	26	0.0009	0.0045	0.0154	0.0107	0.1349	0.0008	< 0.0001	< 0.0001	9.13
P. micra	35977	60	0.0087	0.0047	0.0532	0.4428	0.1040	< 0.0001	< 0.0001	< 0.0001	8.04
F. nucleatum_subsp._vincentii	4538	29	0.0041	0.0048	0.7534	0.1173	0.0150	0.0012	< 0.0001	< 0.0001	2.11
B. wadsworthia	102226	60	0.0070	0.0013	0.0000	0.2000	0.0197	0.0001	< 0.0001	< 0.0001	2.95
A .junii	2735	51	0.0007	0.0002	0.0000	0.3859	0.1411	0.0010	< 0.0001	< 0.0001	8.51
A. rimae	1160	27	0.0012	0.0028	0.0824	0.0525	0.0049	0.0038	0.0020	0.0197	0.06
G. multispecies_spp669_2	43586	95	0.0143	0.0039	0.0000	0.6426	0.3961	0.0035	0.0020	0.0197	5.90
S. gordonii	1470	42	0.0004	0.0004	0.9423	0.4310	0.0884	0.0001	0.0020	0.0197	2.61
S. lactarius	1482	54	0.0004	0.0003	0.0383	0.5216	0.1164	0.0000	0.0020	0.0197	1.40
C. disporicum	261404	88	0.0074	0.0082	0.6160	0.5326	0.2768	0.0055	0.0020	0.0197	4.93
D. pneumosintes	11534	38	0.0012	0.0012	0.9904	0.1523	0.0061	0.0000	0.0020	0.0197	1.71
F. multispecies_spp923_6	1886	24	0.0001	< 0.0001	0.0027	0.0024	0.0001	0.0000	0.0020	0.0197	1.43
F. multispecies_spp930_3	3878	29	0.0003	0.0001	0.0083	0.0934	0.0058	0.0002	0.0020	0.0197	2.37
F. multispecies_spp933_3	75928	63	0.0079	0.0045	0.0532	0.5203	0.1627	0.0001	0.0020	0.0197	2.87
F .multispecies_spp935_4	1898	26	0.0006	0.0001	0.0018	0.0179	0.0003	0.0002	0.0020	0.0197	2.75
K. ascorbata_nov_87.30%	1277	25	0.0020	0.0010	0.1912	0.0851	0.0101	0.0055	0.0020	0.0197	3.87
G. parahaemolysans	10971	33	0.0050	0.0059	0.5264	0.2940	0.0456	0.0006	0.0040	0.0310	0.51
L. fermentum	39573	59	0.0035	0.0070	0.0377	0.0169	0.2266	0.0000	0.0040	0.0310	6.34
S .multispecies_spp386_18	184	38	0.0001	0.0001	0.0023	0.3784	0.0716	0.0007	0.0040	0.0310	1.16
S. multispecies_spp597_2	357	37	0.0001	0.0002	0.3053	0.3223	0.0029	0.0000	0.0040	0.0310	3.30
B. gnavus	112467	62	0.0019	0.0024	0.3974	0.1019	0.3278	0.0018	0.0040	0.0310	1.43
A. variabilis	970	50	0.0008	0.0004	0.0016	0.3482	0.1502	0.0042	0.0040	0.0310	6.60
L. multispecies_spp767_2	73192	57	0.0063	0.0062	0.9679	0.0401	0.2282	0.0001	0.0060	0.0434	5.44
S .multispecies_spp756_2	1332	29	0.0005	0.0009	0.0155	0.2675	0.0595	0.0015	0.0060	0.0434	1.92

*All models are adjusted for age, sex, BMI and log library size. Only bacteria (at species-level) associated with source of samples at $p \leq 0.05$ after correcting for multiple comparisons are shown. Permutation testing accounts for within subject correlation via random intercept.

**Among non-zero samples.

^Adjusted for multiple testing

Full OTU

k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__NA;g__Gemella;s__multispecies_spp670_3

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__gasseri

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__salivarius

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__intermedius

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__multispecies_spp573_2

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae_[XIV];g__Lachnoanaerobaculum;s__saburreum

k__Bacteria;p__Firmicutes;c__Tissierellia;o__Tissierellales;f__Peptoniphilaceae;g__Anaerococcus;s__vaginalis

k__Bacteria;p__Firmicutes;c__Tissierellia;o__Tissierellales;f__Peptoniphilaceae;g__Parvimonas;s__micra

k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium;s__nucleatum_subsp._vincentii

k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Bilophila;s__wadsworthia

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__junii

k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Atopobium;s__rimae

k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__NA;g__Gemella;s__multispecies_spp669_2

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__gordonii

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__lactarius

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium;s__disporicum

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Dialister;s__pneumosintes

k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium;s__multispecies_spp923_6

k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium;s__multispecies_spp930_3

k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium;s__multispecies_spp933_3

k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium;s__multispecies_spp935_4

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Kluyvera;s__ascorbata_nov_87.30%

k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__NA;g__Gemella;s__parahaemolysans

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__fermentum

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__multispecies_spp386_18

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__multispecies_spp597_2

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia;s__gnavus

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__variabilis

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__multispecies_spp767_2

Table A.3: Results from zero-inflated beta regression models comparing bacteria presence/absence and relative abundance across subject disease ICD codes*

Genus	Total read counts	Non-zero samples	Estimated Mean Relative Abundance (μ)**				Estimated Proportion of Presence (P1)				p-value	AIC difference	p-adjusted^
			Control (N=29)	C24 (N=7)	C25 (N=16)	K86.2 (N=6)	Control (N=29)	C24 (N=7)	C25 (N=16)	K86.2 (N=6)			
Porphyromonas	7008	15	0.0006	0.0903	0.0006	0.0062	0.3333	0.1429	0.0588	0.5000	0.0000	28.03	0.0000
Peptoclostridium	4137	14	0.0022	0.0010	0.0927	0.0025	0.3667	0.0000	0.0588	0.3333	0.0004	12.40	0.0273
Acinetobacter	7916	43	0.0238	0.0386	0.0733	0.0583	0.5333	1.0000	0.8824	0.8333	0.0007	11.20	0.0454
Kluyvera	2000	15	0.0015	0.0061	0.0087	0.0082	0.0667	0.2857	0.4706	0.5000	0.0030	7.83	0.1841
Lactobacillus	251585	40	0.1540	0.0801	0.1371	0.1007	0.8667	0.5714	0.2941	0.8333	0.0033	7.54	0.2072
Aggregatibacter	591	18	0.0001	0.0002	0.0007	0.0049	0.3333	0.5714	0.1765	0.1667	0.0034	7.48	0.2120
Ralstonia	153	15	0.0024	0.0012	0.0280	0.0047	0.2667	0.0000	0.2353	0.5000	0.0038	7.23	0.2349
Capnocytophaga	316	12	0.0001	0.0002	0.0010	0.0018	0.1667	0.2857	0.1176	0.5000	0.0061	6.06	0.3768
Enterococcus	28254	29	0.0385	0.0429	0.0508	0.0316	0.5333	1.0000	0.1765	0.5000	0.0065	5.91	0.4004
Clostridium	100517	29	0.0659	0.0519	0.0475	0.0521	0.5667	0.8571	0.1176	0.6667	0.0099	4.85	0.6112
Gemella	7769	24	0.0042	0.0107	0.0146	0.0077	0.2667	0.2857	0.5294	0.8333	0.0127	4.21	0.7852
Prevotella	53918	36	0.0241	0.0334	0.0972	0.0453	0.7000	0.7143	0.3529	0.6667	0.0225	2.72	1.0000
Pseudomonas	64128	31	0.0526	0.0684	0.1324	0.2021	0.3667	0.8571	0.6471	0.5000	0.0248	2.47	1.0000
Raoultella	31688	9	0.0212	0.0449	0.0400	0.0028	0.0333	0.5714	0.2353	0.0000	0.0255	2.40	1.0000
Slackia	14667	36	0.1066	0.0491	0.0858	0.1034	0.8000	0.4286	0.4706	0.1667	0.0259	2.36	1.0000
Selenomonas	438	13	0.0002	0.0003	0.0035	0.0005	0.2000	0.4286	0.1176	0.3333	0.0298	1.98	1.0000
Haemophilus	41752	34	0.0469	0.0496	0.0570	0.0745	0.6667	0.7143	0.2353	0.8333	0.0521	0.48	1.0000
Atopobium	5767	15	0.0015	0.0017	0.0068	0.0097	0.3667	0.2857	0.0588	0.1667	0.0602	0.08	1.0000
Neisseria	2253	11	0.0102	0.0066	0.0010	0.0158	0.2000	0.2857	0.0000	0.5000	0.0760	-0.57	1.0000
Bilophila	10685	13	0.0131	0.1220	0.0316	0.0121	0.2000	0.4286	0.1765	0.1667	0.0807	-0.74	1.0000
Streptococcus	354060	59	0.1089	0.2092	0.2472	0.1273	0.9667	1.0000	1.0000	1.0000	0.0837	-0.85	1.0000
Leptotrichia	15419	21	0.0036	0.0058	0.0133	0.0140	0.3667	0.5714	0.1765	0.5000	0.0854	-0.90	1.0000

** Only bacteria (at genus-level) associated with ICD code (overall) at p≤0.10 prior to correcting for multiple comparisons are shown. Due to missing BMI on two individuals, numbers are based on 58 tissue samples (for comparability to Table 2 these samples were left out).*

***Among non-zero samples.*

^Adjusted for multiple testing.

Full OTU

k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Porphyromonadaceae;g__Porphyromonas

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae;g__Peptoclostridium

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Kluyvera

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Aggregatibacter

k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ralstonia

k__Bacteria;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium

k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Gemellaceae;g__Gemella

k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Prevotellaceae;g__Prevotella

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Raoultella

k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Eggerthellales;f__Eggerthellaceae;g__Slackia

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Selenomonas

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus

k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Atopobiaceae;g__Atopobium

k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Neisseria

k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Bilophila

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus

k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Leptotrichiaceae;g__Leptotrichia

Table A.4: Results from multivariable zero-inflated beta regression models comparing bacteria presence/absence and relative abundance across subject disease ICD codes for RIH samples (excluding NDRI samples)*

Genus	Total read counts	Non-zero samples	Estimated Mean Relative Abundance (μ)**			Estimated Proportion of Presence (P1)			p-value	AIC difference	p-adjusted^
			C24 (N=7)	C25 (N=17)	K86.2 (N=6)	C24 (N=7)	C25 (N=17)	K86.2 (N=6)			
Porphyromonas	6553	5	0.0901	0.0004	0.0060	0.1429	0.0588	0.5000	0.0001	16.83	0.0031
Atopobium	984	4	0.0005	0.0052	0.0081	0.2857	0.0588	0.1667	0.0001	15.20	0.0066
Enterococcus	2370	13	0.0247	0.0301	0.0174	1.0000	0.1765	0.5000	0.0016	9.41	0.0916
Dialister	5111	6	0.0702	0.0109	0.0010	0.1429	0.1765	0.3333	0.0044	7.14	0.2523
clostridium	895	12	0.0098	0.0086	0.0099	0.8571	0.1176	0.6667	0.0053	6.72	0.3036
Stomatobaculum	961	8	0.0006	0.0074	0.0045	0.2857	0.1765	0.5000	0.0083	5.71	0.4717
Neisseria	1982	5	0.0025	0.0004	0.0074	0.2857	0.0000	0.5000	0.0186	3.84	1.0000
Lactobacillus	837	14	0.0018	0.0084	0.0031	0.5714	0.2941	0.8333	0.0190	3.79	1.0000
Ralstonia	100	7	0.0022	0.0292	0.0054	0.0000	0.2353	0.5000	0.0204	3.62	1.0000
Aggregatibacter	569	8	0.0003	0.0008	0.0051	0.5714	0.1765	0.1667	0.0260	3.05	1.0000
Mogibacterium	55	4	0.0001	0.0015	0.0003	0.2857	0.0588	0.1667	0.0269	2.97	1.0000
Capnocytophaga	310	7	0.0002	0.0011	0.0019	0.2857	0.1176	0.5000	0.0278	2.89	1.0000
Propionibacterium	5	5	0.0001	0.0004	0.0001	0.2857	0.0588	0.3333	0.0289	2.80	1.0000
Granulicatella	1649	14	0.0014	0.0023	0.0023	0.7143	0.2353	0.8333	0.0414	1.95	1.0000
Megasphaera	1765	6	0.0025	0.0203	0.0035	0.2857	0.0588	0.5000	0.0457	1.70	1.0000

**Only bacteria (at genus-level) associated with ICD code (overall) at p<0.10 before correcting for multiple comparisons are shown; given small numbers, these models are marginal models for the ICD codes without other covariates. Only Porphyromonas remained statistically significant after adjusting for previous chemotherapy and presence of stent.*

***Among non-zero samples.*

^Adjusted for multiple testing.

Full OTU

k__Bacteria;p__Bacteroidetes;c__Bacteroides;o__Bacteroidales;f__Porphyromonadaceae;g__Porphyromonas

k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Atopobiaceae;g__Atopobium

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Dialister

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Stomatobaculum

k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Neisseria

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus

k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Ralstonia

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Aggregatibacter

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae_[XI];g__Mogibacterium

k__Bacteria;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga

k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae;g__Granulicatella

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Megasphaera

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Selenomonas

k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Eggerthellales;f__Eggerthellaceae;g__Slackia

k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus

k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Enterobacter

Appendix B

Declarations and Supplementary Material for Chapter 2

Declarations

Acknowledgement

We would like to extend our gratitude to Dr. Dong Pei, Lisa Neums, Stefan Graw, Qing Xia, Bo Zhang, Rosalyn Henn and Duncan Rotich of the Department of Biostatistics & Data Science at the University of Kansas Medical Center for their constructive feedback on the methodology.

Ethics approval and consent to participate

The study was approved by Lifespan's Research Protection Office for recruitment at RIH, as well as the Institutional Review Boards for Human Subjects Research at Brown University, Tufts University, and the Forsyth Institute.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. The datasets analyzed for this manuscript will also be available through the NCBI under the BioProject accession no.: PRJNA421501 accompanying the publication of Chung et al. (2019).

R scripts utilized to perform the simulation studies and to analyze the pancreatic cancer dataset are available via the GitHub directory provided in the reference section (Meier, 2019). An example script showing the analysis of a simple pseudo-dataset is also available in the same directory.

Competing interests

K.T. Kelsey is a consultant/advisory board member at Celintec. No potential conflicts of interest were disclosed by the other authors.

Funding

Research reported in this publication was supported by NIH/National Cancer Institute grants R01 CA166150 and P30 CA168524 as well as the the Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, supported in part by the National Institute of General Medical Science award P20GM103428.

Additional File 1

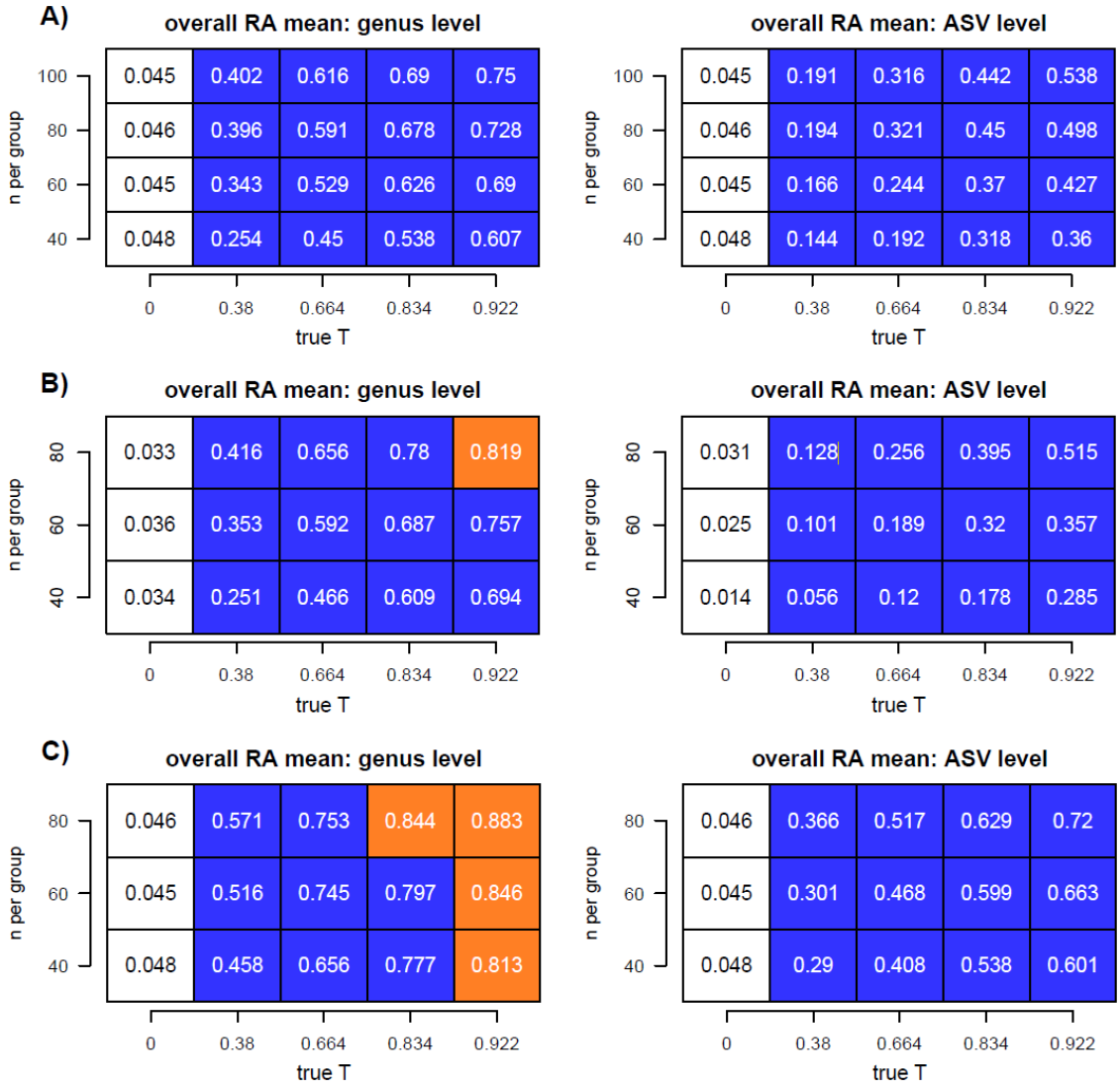


Figure B.1: Additional results of the simulation study for μ. A) depicts the case when calibrating the lower bound of the credible interval of T_θ for a type 1 error rate of 0.05; B) depicts the case when allowing none of the strata to contain exclusively zero valued relative abundances; C) depicts the case when both conditions from A and B are met simultaneously. In part A, H_0 was rejected if $Pr(T_\theta|\mathbf{Y} \leq 0) < 0.05$. In part B and C, H_0 was rejected if $Pr(T_\theta|\mathbf{Y} \leq 0) < q$, where q was adjusted for calibration. Power plots are displayed for testing PASTA of μ with $t_c = 0$ at both ASV and genus level. The term "n per group" refers to the number of samples available in each of the eight sub-group combinations resulting from two body sites and four different levels of disease status. Type 1 error rates are displayed in white colored boxes with black fonts. Power values less than 0.8 are colored blue, values larger than 0.9 are colored red and values between 0.8 and 0.9 are colored orange. While calibration does improve the power compared to the original simulations, restricting sparseness in strata leads to an even stronger improvement in performance.

Additional File 2

Supplementary Simulation Studies

Three additional types of simulations were performed evaluating the PASTA test with respect to the Pearson correlation statistic and the mean response (referred to as θ for convenience) in other modeling settings which we will denote as Scenario A,B and C.

In Scenario A, 8 mean parameters (four θ_{sg} in each site s) were drawn from a chisquared distribution with 10 degrees of freedom (DGOF) in one and with 200 DGOF in another sub-scenario. These numbers were chosen to represen small and large counts (in reference to the rarefying total of 1200 in the ASV data). In each run, means were drawn until $|T_\theta - t| < 0.001$ was satisfied. Means satisfying the condition were subsequently used to draw random samples from Poisson distributions to generate the pseudo response data $Y_k \sim Poisson(\theta_{s_k,g_k})$.

In Scenario B, which aimed to mimic fitting a log-ratio Aitchison model to microbial abundance data, 8 pairs of values (a,b) were drawn. Each value was drawn from a chisquared distribution with 10 DGOF and 200 DGOF respectively in the two considered sub-scenarios. Each pair was used to calculate a mean parameter in the following way: $\theta_{sg} = \log a_{sg} - \log b_{sg}$. In each run, pairs were redrawn until $|T_\theta - t| < 0.001$ was satisfied. Pseudo response data for each observation k was then generated as follows: $Y_k = \log \frac{1+A_k}{1+B_k}$, where A_k was drawn from a $Poisson(a_{s_k,g_k})$ distribution and B_k was drawn from a $Poisson(b_{s_k,g_k})$ distribution.

In Scenario C, which aimed to evaluate a Zero-inflated Poisson (ZIP) model, 8 probabilities of excess absence (p_{sg}) were first drawn from a $Beta(1.67, 0.4)$ distribution (the sampling distribution of ZIBR probabilities of absence on the genus level). Next, 8 Poisson means ($\mu_{s,g}$) were drawn from a chisquared distribution with 10 degrees of freedom (DGOF) in one and with 200 DGOF in another sub-scenario. These values were used to calculate an overall mean parameter in the following way: $\theta_{sg} = (1 - p_{sg}) \cdot \mu_{sg}$. In each run, overall means were drawn until $|T_\theta - t| < 0.001$ was satisfied. Pseudo response data for each observation k was then generated as follows: $Y_k = X_k \cdot Z_k$, where was X_k was drawn from a $Bernoulli(p_{s_k,g_k})$ distribution and Z_k was drawn from a $Poisson(\mu_{s_k,g_k})$

distribution. In order to avoid model convergence issues, new parameters and pseudo datasets in a single run were generated until in each stratum (s,g) there was at least one non-zero observation. In each scenario, a different Bayesian regression model was fit and the PASTA test with respect to θ_{sg} was performed. The fitted models are summarized below:

- Scenario A (Poisson Regression):
  - Likelihood: $Y_k \sim Poisson(\theta_k)$ where $\log(\theta_k) = \beta_{s_k,g_k}$
  - Priors: $\pi(\beta_{s,g}) \sim N(0,100)$
- Scenario B (Normal Regression, Log Count Ratios):
  - Likelihood: $Y_k \sim N(\theta_k, 1/\tau)$ where $\theta_k = \beta_{s_k,g_k}$
  - Priors: $\pi(\beta_{s,g}) \sim N(0,100)$ and $\pi(\tau) \sim Gamma(0.01,0.01)$
- Scenario C (Zero-inflated Poisson Regression):
  - Likelihood: $Y_k \sim f(y_k) = \Big(I(y_k=0)\cdot p_k + I(y_k>0)\cdot(1-p_k)\Big)\cdot \theta_k^{y_k} e^{-\theta_k}/(y_k!)$ where $\log(\theta_k) = \beta_{s_k,g_k}$
  - Priors: $\pi(\beta_{s,g}) \sim N(0,100)$

Results of these additional simulation studies are summarized in **Supplementary Figure 2**. As expected, larger DGOF, i.e. testing based on large counts, consistently led to an increase in statistical power compared to smaller DGOF in all simulation scenarios. Both Scenario A and B were very performant, reaching approximately 0.8 power for the moderate effect of true $T_\theta = 0.66$ and only fitting 10 samples per group. Type 1 error rates appeared to also be adequately calibrated in both cases. Performance metrics were overall very similar between the two scenarios though the simple Poisson regression scenario performed slightly better, likely due having to fit less parameters in its likelihood.

Scenario C performed substantially worse than the two scenarios. Type 1 error rates appeared

mostly calibtrated, though were slightly deflated compared to A and B. In both sub-scenarios adequate power of 0.8 was only achieved when utilizing 60 samples per group and a true $T_\theta = 0.83$. The performance metrics were similar to the calibrated and sparsity restricted simulation evaluating the PASTA test for the overall mean on the genus level in the ZIBR model (row C) in **Supplementary Figure 1**). This makes sense, considering both models use the same sampling distribution for the mixture parameter (p) and both simulations are approximately type 1 error calibrated and subject to the same sparsity restriction.

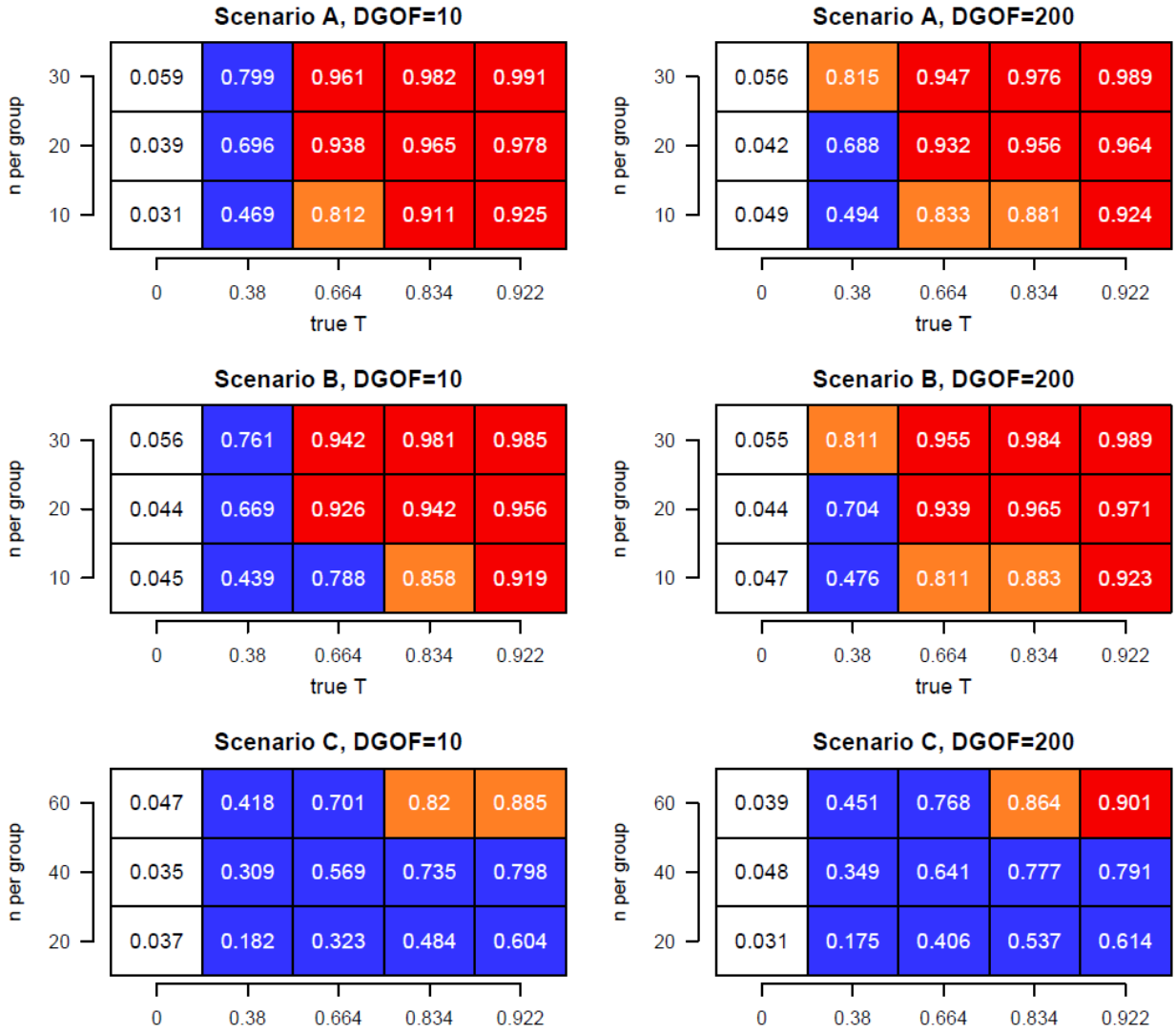


Figure B.2: Results of the supplementary simulation studies. Each row represents one simulation scenario. Scenario A represents a Poisson regression model, Scenario B represents a log count ratio Aitchison model and Scenario C represents a Zero-inflated Poisson regression model. DGOF refers to the degrees of freedom of the chisquared distribution used to sample means of Poisson distributions, that were in turn used to generate count data, used to form pseudo response values. Large DGOF mimic testing highly abundant microbes and small DGOF mimic testing microbes with low abundance. In each scenario, statistical power and type 1 error was evaluated when performing a PASTA test for the mean of the response. The term "n per group" refers to the number of samples available in each of the eight sub-group combinations resulting from two body sites and four different levels of disease status. Type 1 error rates are displayed in white colored boxes with black fonts. Power values less than 0.8 are colored blue, values larger than 0.9 are colored red and values between 0.8 and 0.9 are colored orange.

Additional File 3

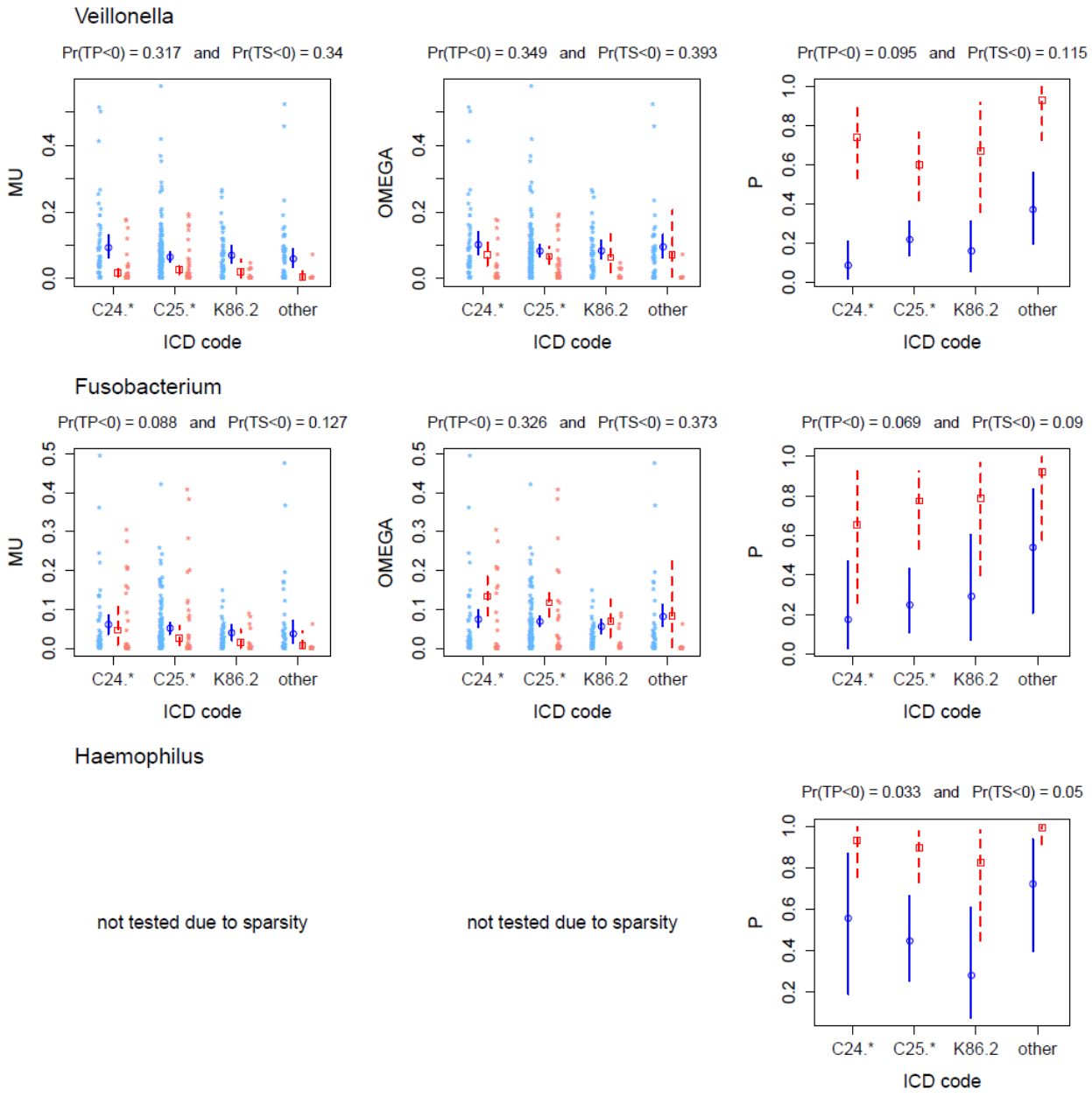


Figure B.3: Plots of parameter estimates within strata when testing for PASTA between gut and mouth on the genus level. Only OTUs with at least marginal significance are displayed. Each row displays the results of an OTU for the three main population parameters of interest. PASTA test results are summarized above each plot. "TP" is T_θ when utilizing Pearson correlation and "TS" is T_θ when utilizing Spearman correlation. Within each plot, circles and squares represent the posterior mean, while vertical lines represent 95% credible intervals. Body site is color coded in red and blue. For μ and ω, relative abundance values are plotted next to the credible intervals.

Appendix C

Declarations and Supplementary Material for Chapter 3

Table C.1: Cell type specific power analysis comparing spatial SCM2 and non-spatial TM models when testing for differential methylation of individual CpGs in simulation C. The rejection rule employs the q% credible interval, where q is chosen such that the type 1 error is controlled at a 10% level. In cases where the widest considered credible interval leads to a larger type 1 error rate in spatial models, both model classes are calibrated to this error rate instead. Here, "alpha" refers to the cell type specific parameter that was used to draw cell proportions via a Dirichlet distribution; a higher value corresponds to consistently drawing higher cell proportions. "t1e" denotes type 1 error rate and "pow" denotes statistical power. The column "pow.diff" containes differences in power obtained when subtracting TM values from SCM2 values.

alpha/cell type	$\sigma^{[B]}$	$\sigma^{[C]}$	t1e.TM	t1e.SCM2	pow.TM	pow.SCM2	pow.diff
6.4/Neutrophil	0.01	0.01	0.10	0.10	>0.99	>0.99	<0.01
6.4/Neutrophil	0.01	0.1	0.42	0.41	0.88	0.90	0.02
6.4/Neutrophil	0.1	0.01	0.10	0.10	0.25	0.37	0.12
6.4/Neutrophil	0.1	0.1	0.10	0.10	0.23	0.30	0.07
3.2/NK	0.01	0.01	0.10	0.10	0.95	0.99	0.04
3.2/NK	0.01	0.1	0.15	0.14	0.56	0.67	0.11
3.2/NK	0.1	0.01	0.10	0.10	0.14	0.18	0.04
3.2/NK	0.1	0.1	0.10	0.10	0.14	0.17	0.03
1.6/Bcell	0.01	0.01	0.10	0.10	0.73	0.86	0.13
1.6/Bcell	0.01	0.1	0.09	0.09	0.38	0.52	0.14
1.6/Bcell	0.1	0.01	0.11	0.11	0.13	0.14	0.01
1.6/Bcell	0.1	0.1	0.11	0.11	0.13	0.14	0.01
0.8/CD4T	0.01	0.01	0.10	0.10	0.67	0.82	0.15
0.8/CD4T	0.01	0.1	0.12	0.12	0.34	0.48	0.14
0.8/CD4T	0.1	0.01	0.10	0.10	0.13	0.16	0.03
0.8/CD4T	0.1	0.1	0.11	0.10	0.13	0.16	0.03
0.4/CD8T	0.01	0.01	0.10	0.10	0.77	0.87	0.10
0.4/CD8T	0.01	0.1	0.22	0.21	0.52	0.59	0.07
0.4/CD8T	0.1	0.01	0.10	0.09	0.14	0.21	0.07
0.4/CD8T	0.1	0.1	0.09	0.09	0.14	0.20	0.06
0.2/Monocyte	0.01	0.01	0.10	0.10	0.86	0.93	0.07
0.2/Monocyte	0.01	0.1	0.23	0.22	0.58	0.65	0.07
0.2/Monocyte	0.1	0.01	0.09	0.09	0.16	0.19	0.03
0.2/Monocyte	0.1	0.1	0.10	0.10	0.16	0.19	0.03

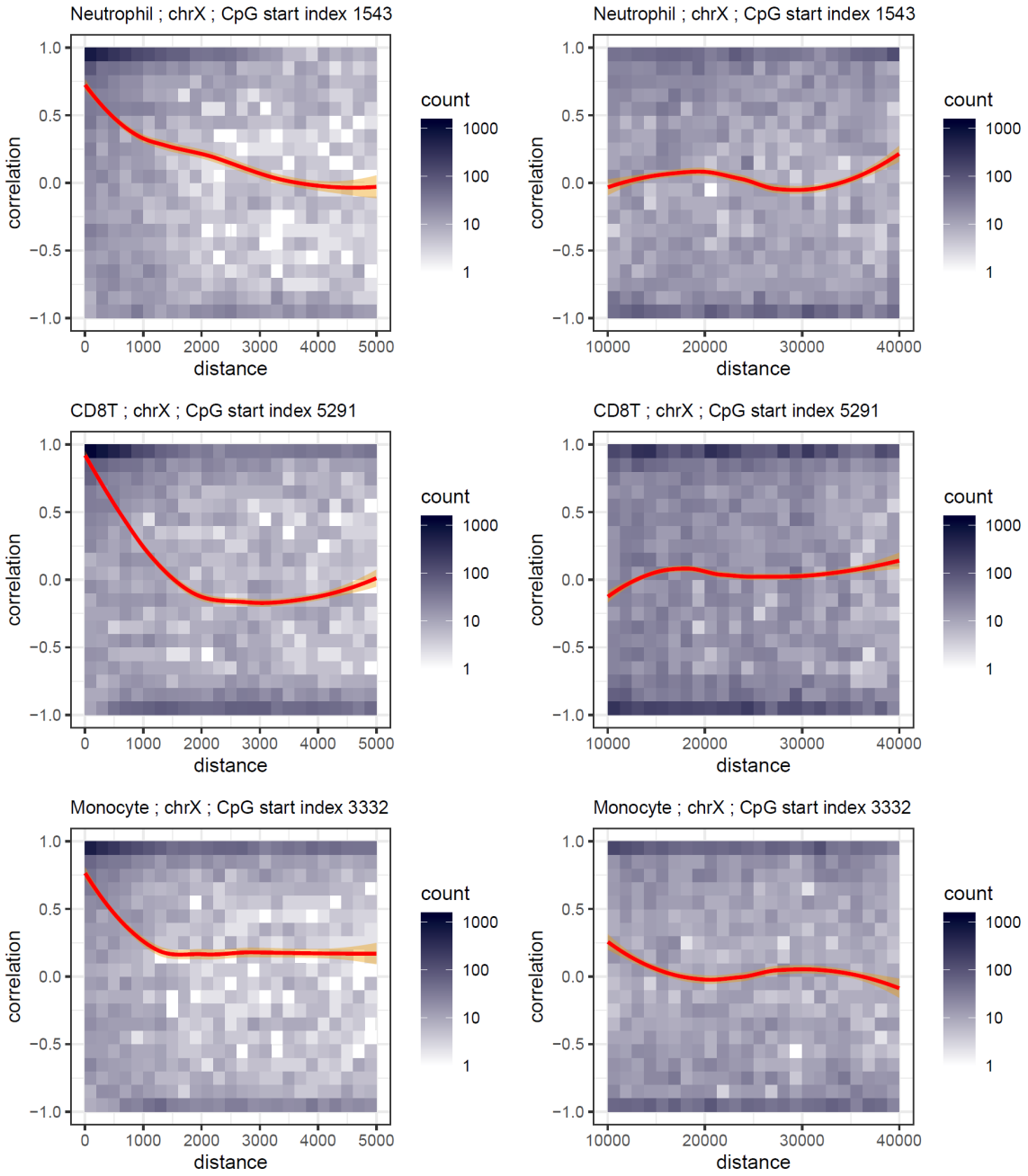


Figure C.1: Pearson correlations of sample beta values for blood cell types as a function of base-pair distance in chromosome X. Orange lines in each plot represent loess smoothed correlation values via the "ggplot2" R-package. While the genomic location was picked at random, the cell types were selected to showcase the variety of trends observed in the data. Overall, smooth trends are similar to those observed in autosomes.