The microbiota of the proximal and distal human colon

Running title: The microbiome of the proximal and distal human colon

Kaitlin J. Flynn[1], Charles C. Koumpouras[1], Mack T. Ruffin IV[2], Danielle Kimberly Turgeon[3†], and Patrick D. Schloss[1†]

† Corresponding authors: kturgeon@med.umich.edu and pschloss@umich.edu

1. Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan 48109

2. Pennslyvania State University, Hershey, Pennslyvania ??

3. Department of Internal Medicine, Division of Gastroenterology, University of Michigan Medical School, Ann Arbor, Michigan

Abstract

Chemical and nutrient gradients along the human colon create microenvironments that affect the distrubution and composition of the gut microbiota. The microbiome has been implicated in colorectal cancer (CRC) and inflammatory bowel disease (IBD). Further, these diseases exhibit different symptoms depending on the location of the colon they are found in. CRC tumors of the proximal and distal colon are morphologically and genetically distinct. Similarly, inflammatory bowel diseases such as Crohn's are typically exacerbated in the proximal instestine while ulcerative colitis patients often experience symptoms in the distal colon. Previous analysis of the fecal microbiota from healthy and CRC or IBD patients has revealed different microbial signatures associated with these diseases. We extended these observations of the fecal microbiome to include analysis of the proximal and distal healthy human colon. We used a two-colonoscope approach on subjects that had not undergone standard bowel preparation procedure. In contrast to previous efforts using prepared colonoscopy, this technique allowed us to characterize the native proximal and distal luminal and mucosal microbiome without prior chemical disruption. 16S rRNA gene sequencing was performed on proximal and distal mucosal and luminal biopsies and home-collected stool for 20 healthy individuals. Diversity analysis revealed that each site contained a diverse community and that a patient's samples were more similar to each other than to that of other individuals. A patient's feces were most similar to samples taken from the distal lumen, likely reflecting the anatomical structure of the colon. Since we could not differentiate the communities in the proximal and distal colon based on community structure or community membership alone, we employed the Random Forest machine-learning algorithm to identify key taxa that the two sites in the lumen and mucosa. Random Forest classification models were built using taxa abundance and sample location and revealed distinct populations that were found in each location. A model differentiating the proximal mucosa and lumen was built with an AUC of 0.856. The proximal mucosa had a higher abundance of the genera *Enterobacteriaceae* and *Bacteriodes*. *Peptoniphilus, Anaerococcus, Finegoldia,* and *Turicibacter* were most likely to be found in distal mucosal samples versus distal luminal samples (AUC = 0.975). The classification model performed well (AUC = 0.920) when classifying mucosal samples into proximal or distal sides, but separating luminal samples from each side proved more challenging (AUC = 0.575). The distal mucosa was found to have high populations of *Finegoldia, Murdochiella* and *Porphyromonas*. Proximal and distal luminal samples were comprised of many of the same taxa, likely reflecting the fact that stool moves along the colon from

the proximal to distal end. By sampling the unprepped human colon, our results have identified distinct bacterial populations native to the proximal and distal sides. Further investigation of these bacteria may elucidate if and how these groups contribute to different pathogenesis processes on the respective sides of the colon.

wordcount?

Introduction

The human colon is an ecosystem comprised of several different microenvironments inhabited by resident bacterial members of the microbiome. Concentrations of oxygen, water and anti-microbial peptides change along the gut axis and influence what populations of microbes reside in each location. Microenvironments differ not only longitutinally along the colon, but latitudinally from the epithelium to mucosa to intestinal lumen, offering several sites for different microbial communities to flourish. The identity of these specific microbes and communities are important for understanding the etiology of complex colon diseases such as Inflammatory Bowel Disease (IBD) and Colorectal Cancer (CRC). IBD and CRC are known to be preceded or accelerated by perturbations in gut microbes (1–3). The severity, symptoms, morbidity and mortality of these diseases is known to vary based upon the biogeographical location in which they occur. For instance, CRC tumors that arise on the distal side of the colon are infiltrating lesions that present with painful symptoms (4). In contrast, 47% of CRCs are caused by proximal-sided colon tumors that are sessile and form along the wall of the colon, often remaining asymptomatic until advanced carcinogenesis (4). The distal and proximal sides of the colon differ in the amount of inflammation present and the genomic instability of precancerous cells, respectively, in addition to variation in the previously mentioned chemical gradients (1, 5, 6). In IBD patients, disease flares in the distal colon are usually indicative of ulcerative colitis (UC), whereas Crohn's disease (CD) patients typically experience disease in the small intestine, ileum and proximal colon (2). UC presents as large and highly inflammed mucosal ulcers, where as CD lesions are often smaller and have areas of normal tissue distributed amongst the flare (2). Thus, given the varied physiology of the proximal-distal axis of the colon and known differences in disease patterns at these sites, symbiotic microbes and their metabolites likely vary as well, and may influence the heterogenous disease prognoses of IBD and CRC. Because CRC can be a long-term complication of

IBD, the distribution of microbes is important to understanding the pathophysiology of both diseases.

Several recent findings have shown that development and progression of IBD or CRC can be attributed to specific molecular events as a result of interactions between the gut microbiota and human host (1, 3). For instance, comparison of the bacteria present on CRC tumors with those found on nearby healthy tissue has identified specific species that are tumor-associated (7). These species include the oral pathogens *Fusobacterium nucleatum* and *Porphyromonas asacharolytica*. Interestingly, these periodontal pathogens have been highly predictive of whether a patient had CRC tumors or not in our prior human stool classification studies (8). *F. nucleatum* has also been found to be elevated in the stool and biopsies of patients with IBD as compared to healthy controls (9). Furthermore, studies of *F. nucleatum* isolated from mucosal biopsies showed that more invasive *F. nucleatum* positively correlates with IBD disease level (9). Like many intestinal pathogens, the bacteria appear to have a high-impact despite being lowly-abundant in the community (2). The physiology of these rare taxa may contribute to the colonic disease state. These studies often examined only shed human stool or the small intestine, preventing fine-resolution analysis of paired samples from the proximal and distal sides of the colon. Similarly, comparisons of on- or off-tumor/lesion bacteria rarely have matched tissue from the other side of the colon from the same patient, limiting what conclusions can be drawn about the colonic microbiome overall, let alone at that specific site. Due to these limitations, the contribution of the gut microbiota to IBD and CRC disease location in the colon is largely undefined. Characterizing these communities could provide needed insight into disease etiology, including how the disruption of the healthy community could promote the initiation or proliferation of the distinct proximal and distal CRC tumors or IBD flares.

The few existing profiles of the microbial biogeography of the gut have been limited by sample collection methods. The majority of human gut microbiome studies have been performed on whole shed feces or on samples collected during colonoscopy procedures. While the latter method allows investigators to acquire samples from inside the human colon, typically this procedure is preceded by the use of bowel preparation methods such as the consumption of laxatives to cleanse the bowel. Bowel preparation is essential for detecting cancerous or precancerous lesions in the colon, but complicates microbiome profiling as the chemicals strip the bowel of contents and disrupt the mucosal layer (10, 11). As such, what little information we do have about the biogeographical distribution of the microbes in the proximal and distal colon is confounded by the bowel preparation procedure.

Here we address the limitations of previous studies and effectively identify the microbes specific to the lumen and mucosa of the proximal and distal healthy human colon. We used an unprepared colonoscopy technique to sample the natural community of each location of the gut without prior disruption of the native bacteria in 20 healthy volunteers. To address the inherent inter-individual variation in human microbiomes, we used a machine-learning classification algorithm trained on curated 16S rRNA sequencing reads to identify microbes specific to each location. We found that our classification models were able to separate mucosal and lumenal samples as well as differentiate between sides of the colon based on populations of specific microbes. By identifying the specific microbes we are poised to ask if and how the presence or disruption of the microbes at each site contribute to the development of the specific tumor subtypes of CRC in the proximal and distal human colon.

Results

Microbial membership and diversity of the proximal and distal colon

Lumenal and mucosal samples were collected from the proximal and distal colon of 20 healthy humans that had not undergone bowel preparation (Figure 1). Participants also collected stool at home one week prior to the procedure. To characterize the bacterial communities present at these sites, 16S rRNA gene sequencing was performed on extracted DNA from each sample. As expected, each site was primarily dominated by *Firmicutes* and *Bacteriodetes* (Figure 2A) (12). Samples had varying levels of diversity at each site, irrespective of the individual (Figure 2B). For example, the proximal mucosa was more diverse than the distal for some individuals while the opposite was true for others. Therefore we could not identify a clear pattern of changes in microbial diversity along the gut axis.

To compare similarity between sides (proximal or distal) or sites (lumen or mucosa), we calculated distances from Operational Taxonomic Unit (OTU) abundances and compared these distances for all individuals. Again, across all patient samples we observed a range of distances when comparing sample locations (Figure 3A) and again those ranges did not follow a clear pattern on an individual basis. However, when comparing median distances between the proximal lumen and mucosa, the proximal versus distal lumen, the proximal versus distal mucosa, and the distal lumen and mucosa, we found that the proximal lumen and mucosa were most similar to each other than the other samples ($P < 0.005$,

Wilcoxon, BH adjustment).

Fecal samples resemble lumenal samples from the distal colon

Next, we calculated distances to examine how each sample compared to the home-collected feces. Amidst variability between patients, we did identify significantly smaller distance between the distal lumenal sample and the feces (Figure 3B, $P < 0.05$, Wilcoxon, BH adjustment). Furthermore, there was an even larger difference in the comparisons of the distal mucosa to the feces, indicating that the mucosa is different from the stool as compared to lumen ($P < 0.0005$, Wilcoxon, BH adjustment). To determine what factors may be driving the differences seen among the samples, we compared distances between samples from all subjects (interpersonal) versus samples from within one subject (intrapersonal). We found that samples from one individual were far more similar to each other than to other study subjects (Figure 3C), consistent with previous human microbiome studies that have sampled multiple sites of the human colon (13–15). Thus interpersonal variation between subjects drove the differences between samples more than sample site or location. Overall, the results comparing the structure of the communities suggest that the contents of the distal lumen are most representative of the patient's feces, and the microbes remaining on the mucosa are more distinct.

Random Forest classification models identify important Operational Taxonomic Units (OTUs) on each side

To identify OTUs that were distinct at each biogeographical site, we constructed several Random Forest models trained using OTU abundances. We used 10-fold cross validation to build the first model to classify the lumenal versus mucosal samples for the proximal and distal sides, independently (Figure S1A). The models were constructed to only use the five most predictive OTUs as input to reduce overfitting between samples. The models performed well when classifying these samples (AUC = 0.856 and AUC = 0.975, respectively). The OTUs that were most predictive of each site were identified by their greatest mean decrease in accuracy when removed from the model. For distinguishing the proximal lumen and mucosa, OTUs from the *Bacteriodes*, *Actinomyces*, *Psuedomonas* and two OTUs from the *Enterobacteraceae* genera were differentially abundant (Figure 4A). The model classifying the distal lumen and mucosa identified OTUs from *Turicibacter*, *Finegoldia*, *Peptoniphilus* and two OTUs from the

Anaerococcus genera (Figure 4B). These results indicated that there were fine differences between the different sites of the colon, and that these could be traced to specific OTUs on each side.

Next, we built a model to differentiate the proximal and distal lumenal samples using 10-fold cross validation. The model performed best when distinguishing the proximal versus distal mucosa (Figure S1B, AUC = 0.920) compared to the proximal versus distal lumen (AUC = 0.575). These models were also constructed to use only the five most predictive OTUs as input to reduce overfitting. OTUs that were differentially abundant between the distal and proximal mucosa included members of the *Porphyromonas*, *Murdochiella*, *Finegoldia*, *Anaerococcus* and *Peptoniphilus* genera (Figure 5A). Differentially abundant OTUs of the proximal and distal lumen included three OTUs of the *Bacteroides* genus, a *Clostridium IV* OTU and an *Oscillibacter* OTU (Figure 5B). This analysis found that some of the same OTUs that are distinct between the mucosa and lumen also helped to differentiate between the two sides- such as *Anaerococcus* and *Finegoldia*.

Bacterial OTUs associated with cancer are found in healthy individuals

Given that specific bacterial species have been associated with colorectal cancer and IBD, we probed our sample set for these OTUs. Among our 100 samples, the most frequent sequence associated with the *Fusobacterium* genus was OTU179, which aligns via BLASTn to *Fusobacterium nucleatum subsp animalis* (100% over full length). This is the only species of *Fusobacterium* known to have oncogenic properties and be found on the surfaces of colorectal cancer tumors. (16). There were 14 patient samples with the *F. nucleatum subsp. animalis* sequences. Of the samples with the highest abundance of *F. nucleatum subsp. animalis*, four of the samples were from the proximal mucosa and three from the distal mucosa (Figure 6A). The second most frequent *Fusobacterium* sequence was OTU472, which aligned with 99% identity to *F. varium*. In addition to *F. nucleatum*, *F. varium* has been associated with IBD (17). Four study participants harbored *F. varium* and the samples were split evenly between the proximal and distal mucosa (Figure 6B). OTU152 was similar to the members of the *Porphyromonas* genus and the most frequent sequence in that OTU aligned to *Porphyromonas asacharolytica* (99% over full length), another bacterium commonly detected and isolated from colorectal tumors. OTU152 was only detected on the distal mucosa, and in fact was one of the OTUs the classification model identified as separating distal and proximal sides (Figure 6C). Among the 11 distal mucosa samples that were

positive for *P. asacharolytica*, the relative abundances for this OTU ranged from 0.01% - 16%. Thus, disease-associated OTUs could be found in our sample set of 20 healthy individuals.

Discussion

Here we identified bacterial taxa that were specific to the lumen and mucosa of the proximal and distal sides of the human colon from samples collected during unprepared colonoscopy. We found that all locations contained a range of phyla abundances and a range of diversity, but that there was a wide variability between subjects. Pairwise comparisons of each of the sites revealed that the proximal mucosa and lumen were most similar to each other. Further, comparison of colonoscopy-collected samples with samples collected from stool at home showed that the distal lumen was most similar to feces. Random Forest models built on OTU relative abundances from each sample identified microbes that were particular to each location of the colon. Finally, we were able to detect some bacterial OTUs associated with colonic disease in our healthy patient cohort. Using unprepped colonoscopy and machine learning, we have identified bacterial taxa specific to the healthy proximal and distal human colon.

When examining the relative abundance of the dominant phyla at each site (i.e. *Bacteriodes* and *Firmicutes*), there was a wide amount of variation. This likely reflects not only the variability between human subjects, casued by differences in age, sex, diet, but also biogeogrphical "patchiness" in the gut microbiome. Several studies have noted that the bacteria recoverable from the same mucosal sample location can be vastly different when the samples are taken just 1 cm away from each other (18). Similar patchiness is also observed in lumenal contents and fecal samples themselves; there is observed separation of different interacting microbes along the length of a stool sample, for instance (19). That said, across our samples the mucosal samples harbor more *Proteobacteria*, consistent with previous studies comparing mucosal swabs to lumenal content in humans (5). Hence, the conclusions we can draw from phyla analysis may be impacted by patchiness between subjects.

To get around the noisiness from a diverse set of samples, we built a Random Forest model to identify microbes specific to each side. For each comparison we identified the top five OTUs that were strongly predictive of one site or another. Generally, OTUs identified in each location were consistent with known physiological gradients along the gut axis (6). For instance, the proximal mucosa contains the

highest oxygen concentratons of the colon and harbored mucosa-associated facultative anaerobes such as *Actinomyces* and *Enterobacteraceae* and aerobic *Psuedomonas*. The distal mucosa was far more likely to host strictly anaerobic species such as *Porphyromonas*, *Anaerococcus*, *Finegoldia* and *Peptoniphilus*. Thus the gut microenvironment of each location likely enriches for these specific microbes.

In addition to identifying features that are specific to each side of the gut, the ability of the Random Forest to classify samples can serve as a proxy for similarity. That is, a higher AUC value means the samples are more efficiently classfied (and thus more different) than a model with a lower AUC value. For instance, the model separating the proximal and distal mucosa has an AUC of 0.925 whereas the model for classifying the proximal and distal lumen has a much lower AUC of 0.575. Further, prior to reducing the input to five OTUs, the latter model required 44 OTUs to best separate the samples (AUC $= 0.755$). The much lower AUC and need for a high number of features compared to other models suggest these locations are the most similar of the comparisons tested. We speculate that the model was least effective at classifying the proximal and distal lumenal contents because the samples are arguably composed of the same bacteria but differ in water content.

We detected *F. nucleatum* and *P. asacharolytica* in 8 and 5 of our subjects, respectively. These bacteria have been shown to be predictive of colorectal cancer in humans (8) and have oncogenic properties in cell culture and in mice (20). Interestingly, while *F. nucleatum* was found on both sides of the colon, *P. asacharolytica* was only detected in the distal mucosa. Not much is known about the distribution of *P. asacharolytica* but given its documented anaerobic characteristics and asacharolytic metabolism, it might not be surprising that it resides in the less-oygen-rich and proteinaceous distal mucosa (5). In studies examining bacteria on colorectal cancer tumors, *F. nucleatum* is more commonly detected on proximal-sided tumors, and distribution of *F. nucleatum* decreases along the colon to rectum (21). Of the 8 (40%) individuals positive for *F. nucleatum* in this study, the bacterium was spread across the proximal mucosa, distal lumen and distal mucosa. Data examining bacterial biofilms on the mucosa of CRC tumors suggests that *Fusobacteria* species are more commonly found on proximal tumors and in biofilms, indicating that it is not only the presence of the bacteria but the organization of the tumor community that contributes to *Fusobacterium's* role in tumorigenesis (7). Finally, *Fusobacterium* and *Porphyromonas* species have been known to not only co-occur on CRC tumors but also to synergistically promote tumorigenesis in an oral cancer model (22) (23). Thus, further analysis of the distribution and

activities of these pathogens may elucidate a mechanism for development of IBD or CRC subtypes in the proximal or distal colon.

The *Fusobacterium* species *nucleatum* and *varium* have been commonly isolated from mucosal biopsies of patients with IBD (17). Laboratory experiments with these isolates have shown that disease-isolated *F. nucleatum* are more invasive and stimulate more TNF-α production than strains from healthy individuals (9), suggesting the bacteria may increase inflammation in the gut as well (24). *F. varium* isolated from UC patients caused colonic ulcers in an experimental mouse model (25). *F. varium* was only detected in three of our study participants and two of those samples were isolated form the proximal mucosa (Figure 6B). *F. varium* is most commonly isolated from UC patient biopsies from the ileum or cecum (26), suggesting this species may exhibit preference for the different environmental conditions of these gastrointestinal sites. Further work will assess how gut environment may select for species which may then cause localized disease.

Specific comparisons of our findings to previously published gut biogeography studies are confounded by the use of bowel preparation methods in most other studies. A rare report of a matched-colonoscopy study sampled 18 patient's colonic mucosa and lumenal contents prior to and after bowel cleansing (27). This study found that mucosal and lumenal samples were distinguishable prior to bowel cleansing, but that bowel preparation resulted in an increase in shared OTUs between each site (27). After seven days, bowel cleansing not only made the samples harder to distinguish, but it also decreased in diversity across sites. Bowel preparation clearly biases into the microbes recovered from sampling the lumen or mucosa of a prepared bowel.

By revealing specific differences in microbial populations at each location in the gut via sampling an unprepared bowel, we can begin to form hypothesies about how specific host-microbe interactions can affect disease progression of proximal and distal CRC and IBD subtypes. To this point, 16S rRNA gene sequecing community profiling studies do not provide enough information to fully probe these questions. In particular, 16S sequencing cannot not profile the host characteristics at each site. Combining the unprepared colonoscopy approach with analysis of multi-omic sequencing data may be useful in further characterizing host-microbiome interactions along the gut axis for both health and disease.

Acknowledgments

We thank all the individuals who volunteered for the study. This work was supported by the Rose and Lawrence C. Page Foundation (DKT). We would also like to thank Brian Kleiner, Chelsea Crofoot, and Kirk Herman for their roles in study coordination, subject recruitment, sample collection and sample processing.

Methods

Human subjects

The procedures in this study and consent were approved by the Institutional Review Board at the University of Michigan Health System with protocol number XXXX. Subjects were recruited using the online recruitment platform and were pre-screened prior to enrollment in the study. Exclusion criteria included: use of asprin or NSAIDs within 7 days, use of antibiotics within 3 months, current use of anticoagulants, known allergies to Fentanyl or Benadryl, prior history of colon disease, diabetes, abdominal surgery, respiratory, liver, kidney or brain impairments, undergoing current chemotherapy or radiation treatment and subjects that were pregnant or trying to conceive. 20 subjects that met the criteria were selected and provided signed informed consent prior to the procedure. There were 13 female and 7 male subjects ranging in age from 25 to 64.

Sample collection

At a baseline visit, subjects were consented and given a home collection stool kit (Source of kit?). At least one week prior to the scheduled colonoscopy, subjects were to collect whole stool at home and ship the samples to a research coordinator on ice. Notably, subjects did not undergo any bowel preparation method prior to sampling. On the procedure day, subjects reported to the Michigan Clinical Research Unit at the University of Michigan Health System. Patients were consciously sedated using Fentanyl, Versed and/or Benadryl as appropriate. A flexible sigmoidoscope was first inserted about 25cm into the colon and endoscopy brush used to collect lumenal/stool contents. Two lumenal samples were collected and the contents immediately deposited into RNAlater (Fischer) and flash-frozen in liquid nitrogen. The brushes were withdrawn and biopsy forceps were used to collect mucosal biopsies on sections of the

colon that were pink and free of stool matter. Three mucosal biopsies were collected and flash-frozen in RNAlater. These samples comprised the distal or distal colon samples. The sigmoidoscope was then withdrawn and a pediatric colonoscope was inserted to reach the ascending colon. Samples were then collected as in the distal colon and the colonoscope withdrawn. All samples were stored at -80°C.

Sample processing, sequencing and analysis

DNA extraction was performed using the PowerMicrobiome DNA/RNA Isolation Kit (MO BIO Laboratories). For tissue biopsies, Bond-Breaker TCEP solution (Fisher) and 2.8mm ceramic beads (MO BIO Laboratories) were added to the bead beating step to enhance DNA recovery from mucosal samples. The resulting DNA was used as template for amplification of the V4 region of the 16S rRNA gene and fragments were sequenced on an Illumina MiSeq as previously described (28). Sequences were curated using the mothur software as described previously (29). The sequences were assigned taxonomic classification using a naive Bayesian classifier trained using a 16S rRNA gene training set from the Ribosomal Database Project (RDP) (30) and clustered into operational taxonomic units (OTUs) based on a 97% similarity cutoff. Sequencing and analysis of a mock community revealed the error rate to be 0.018%. Samples were rarefied to 4231 sequences per sample in order to reduce uneven sampling bias.

Diversity analysis was performed using the Simpson diversity calculator and θYC calculator metrics in mothur version 1.39.5 (29). ThetaYC distances were calculated to determine the dissimilarity between two samples. Random Forest classification models were built using the randomForest R package and resultant models were used to identify the OTUs that were most important for classifying each location (31). To get species-level information about sequences of interest, sequences were aligned using blastn and the species name was only used if the identity score was $\geq$ 99% over the full-length of the contig and matched a single reference.

Statistical analysis

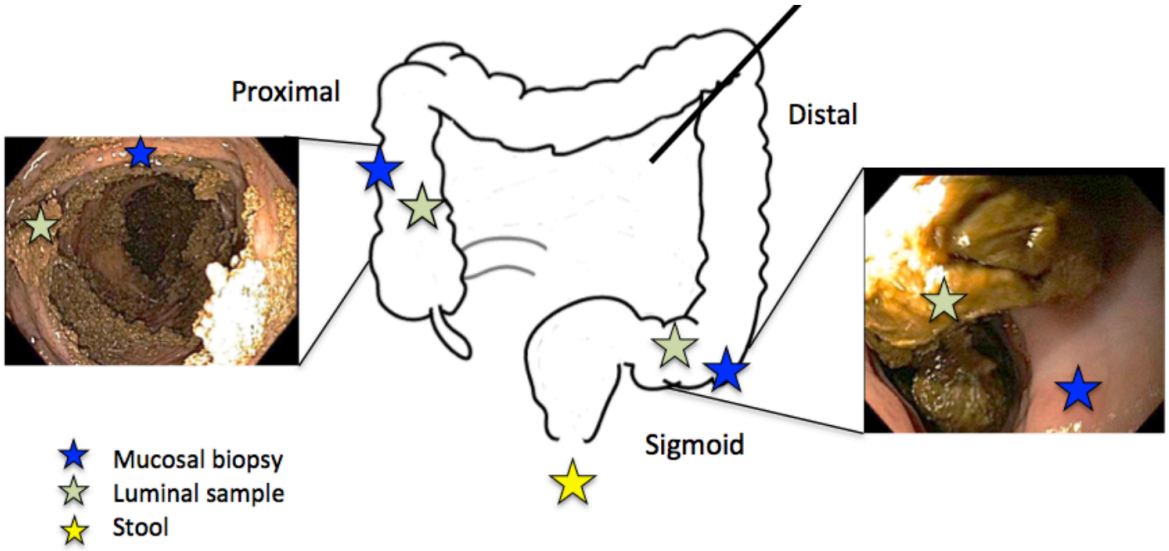
Differences in community membership at the phyla level were tested using the analysis of molecular variance (AMOVA) metric in mothur. Differences in θYC distances by location were tested using the Wilcoxon rank-sum test adjusted for multiple comparisons using the Benjamini-Hochberg procedure.

Data availability

16S rRNA gene sequence reads and experiment metadata are available on the NCBI Sequence Read Archive (SRA) with accession number XXXX. A reproducible data analysis pipeline can be found at https://github.com/SchlossLab/Flynn_LRColon_XXXX_2017.

Figures

Figure 1



Sampling strategy. A flexible sigmoidoscope was used to sample the distal colonic luminal contents and mucosa. The scope was inserted ~ 25cm into the subject and endoscopy brushes were used to sample the luminal contents (green star). A separate set of biopsy forceps was used to sample the distal mucosa (blue star). The sigmoidoscope was removed. A pediatric colonoscope was inserted and used to access the proximal colon. Biopsies were taken of the proximal luminal contents and mucosa as described. One week prior to the procedure stool was collected at home and sent into the laboratory. Representative images from one individual are shown.

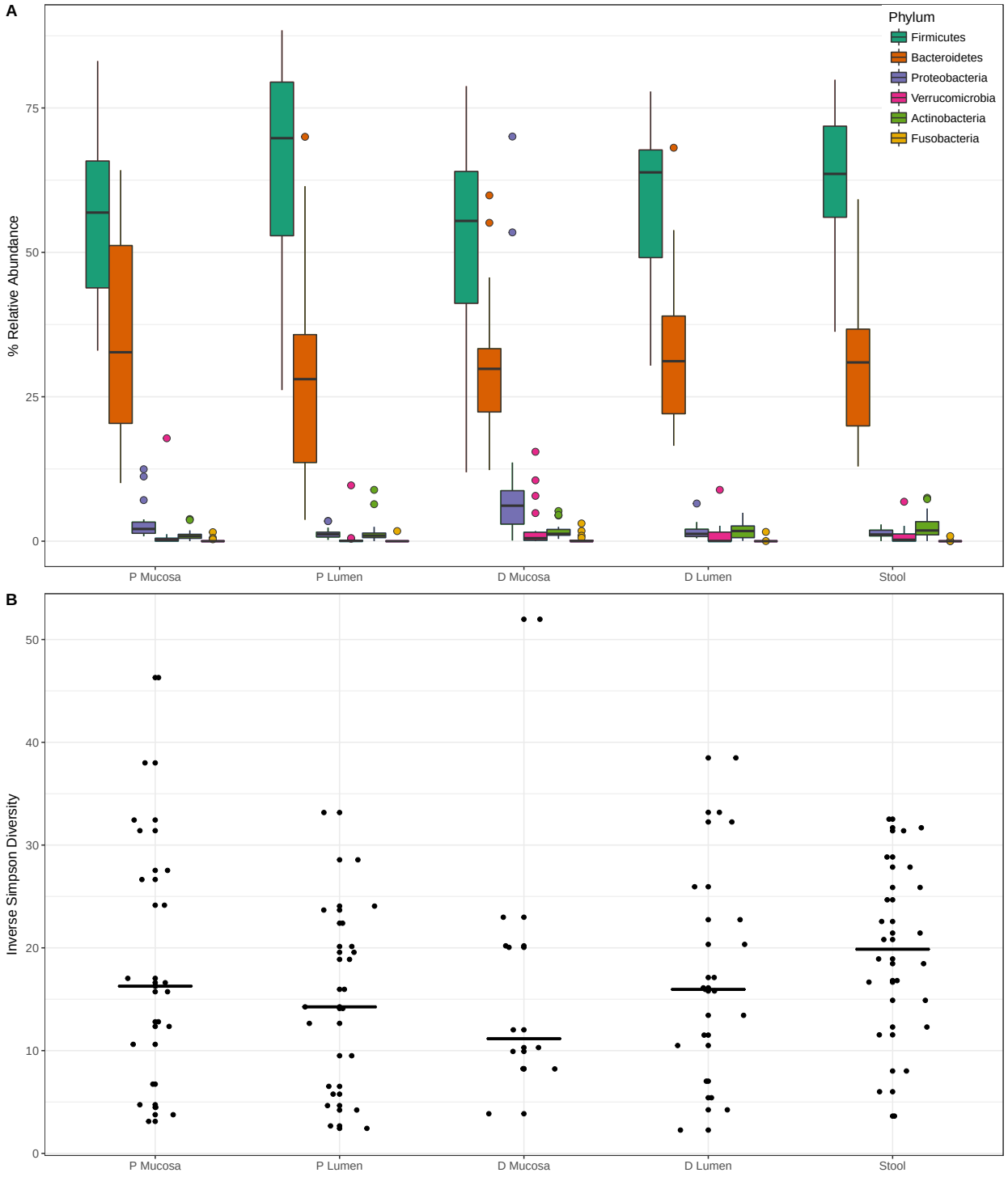
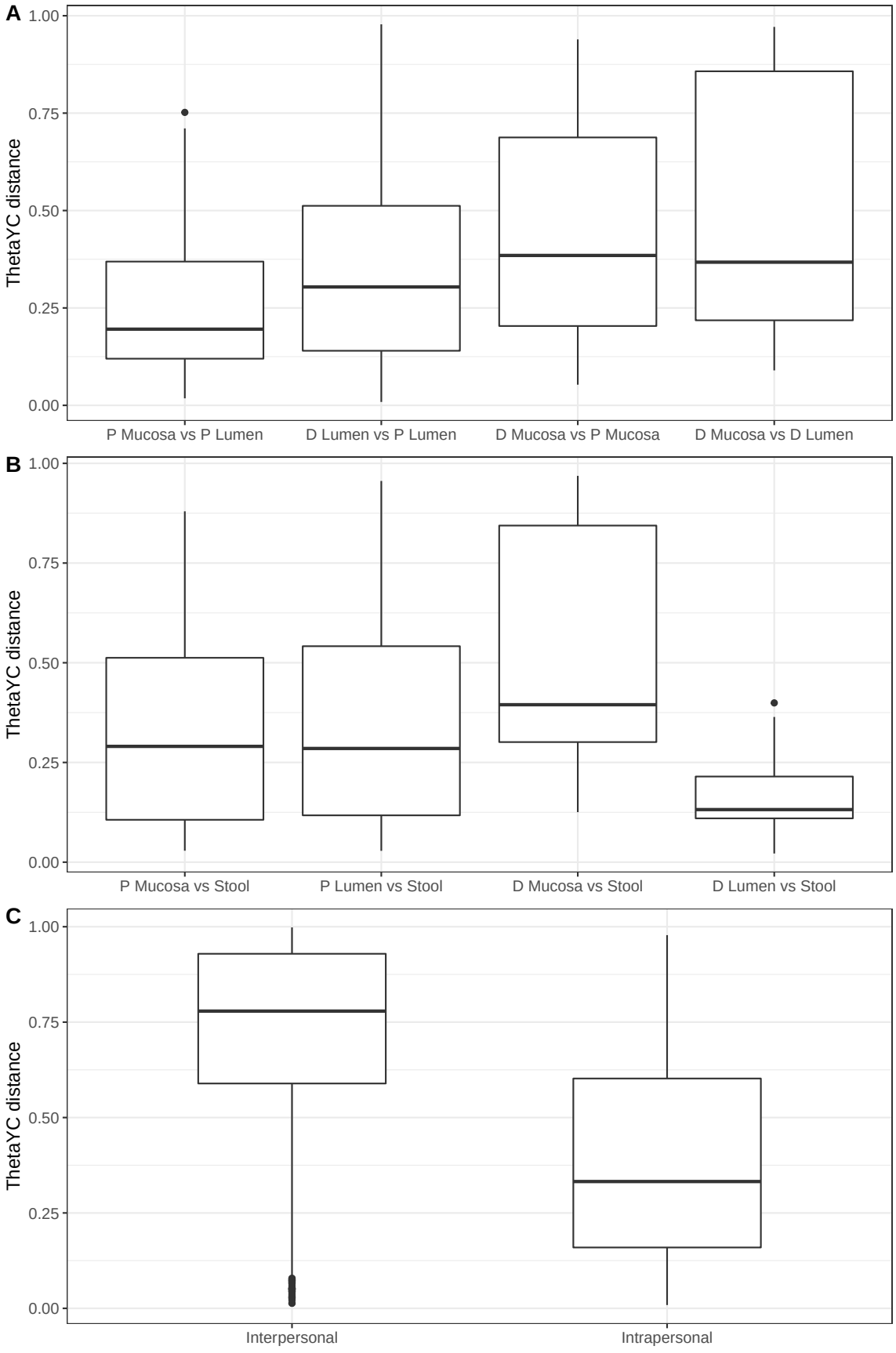


Figure 2

Microbial membership and diversity of the proximal and distal human colon. A) Relative abundance of the top five bacterial phyla in each sampling site. Each box represents the median and confidence

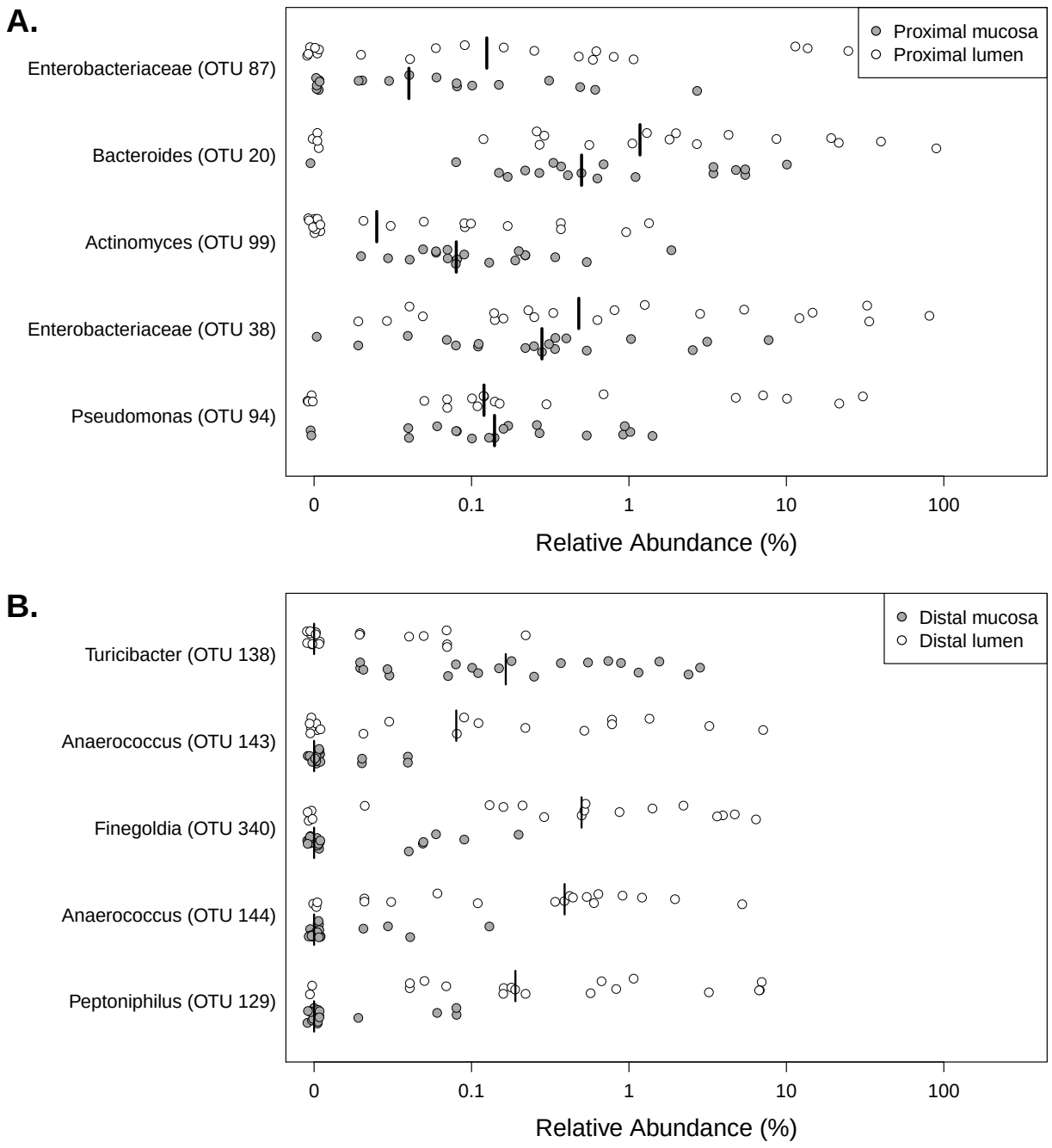
intervals. B) Simpson diversity of the microbial communities at each location. The lines represent the median values.


A
1.00
0.75
0.50
0.25
0.00
ThetaYC distance
P Mucosa vs P Lumen
D Lumen vs P Lumen
D Mucosa vs P Mucosa
D Mucosa vs D Lumen
B
1.00
0.75
0.50
0.25
0.00
ThetaYC distance
P Mucosa vs Stool
P Lumen vs Stool
D Mucosa vs Stool
D Lumen vs Stool
C
1.00
0.75
0.50
0.25
0.00
ThetaYC distance
Interpersonal
Intrapersonal

333

Figure 3

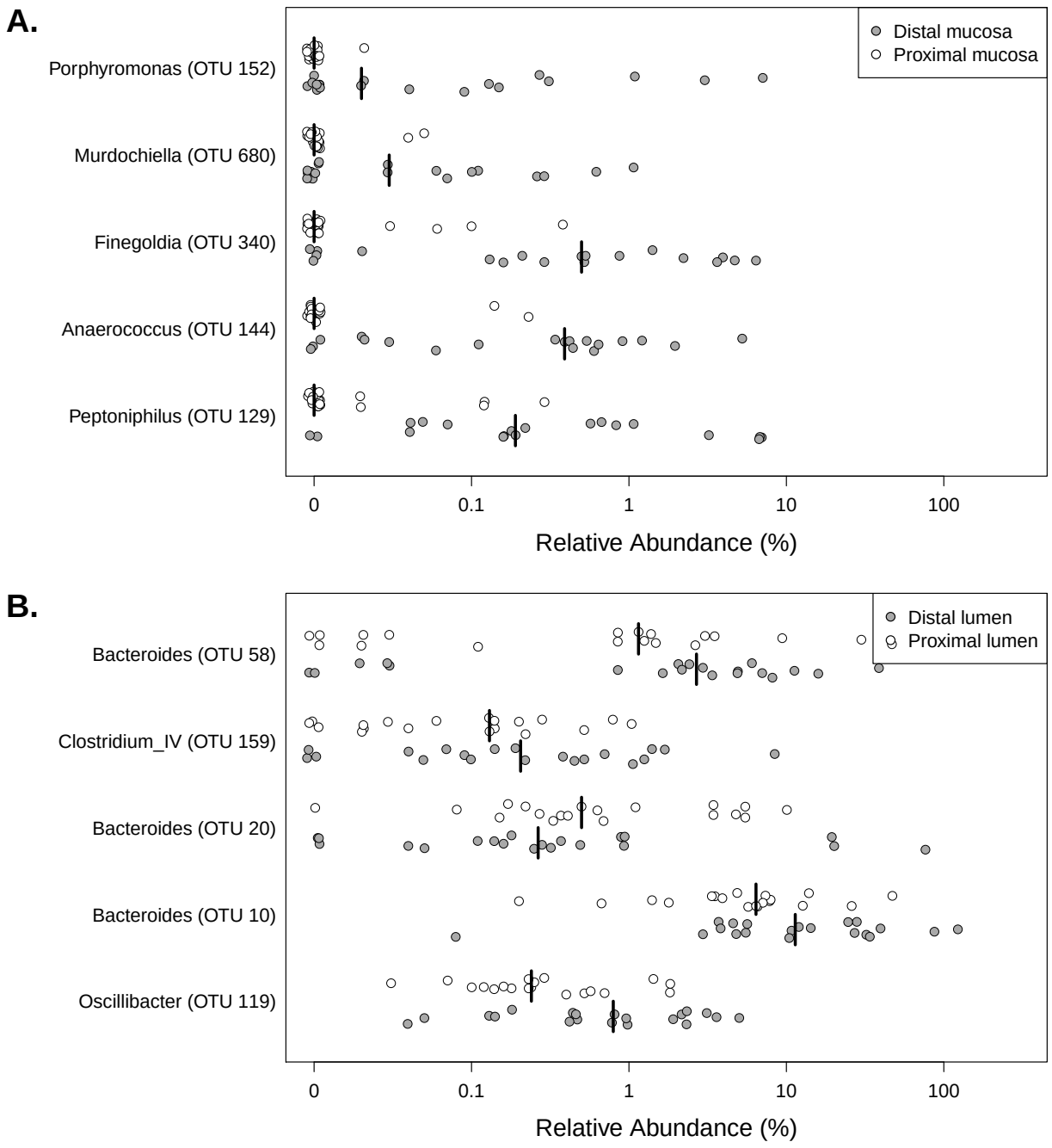
Distances of microbial community structure between sites of the gut. ThetaYC distances are shown for interpersonal similarities between two sites – each point represents one individual. In (A), comparisons of the proximal and distal mucosal and lumen are shown. In (B), comparisons of each site to the exit stool are shown.



339

340 **Figure 4**

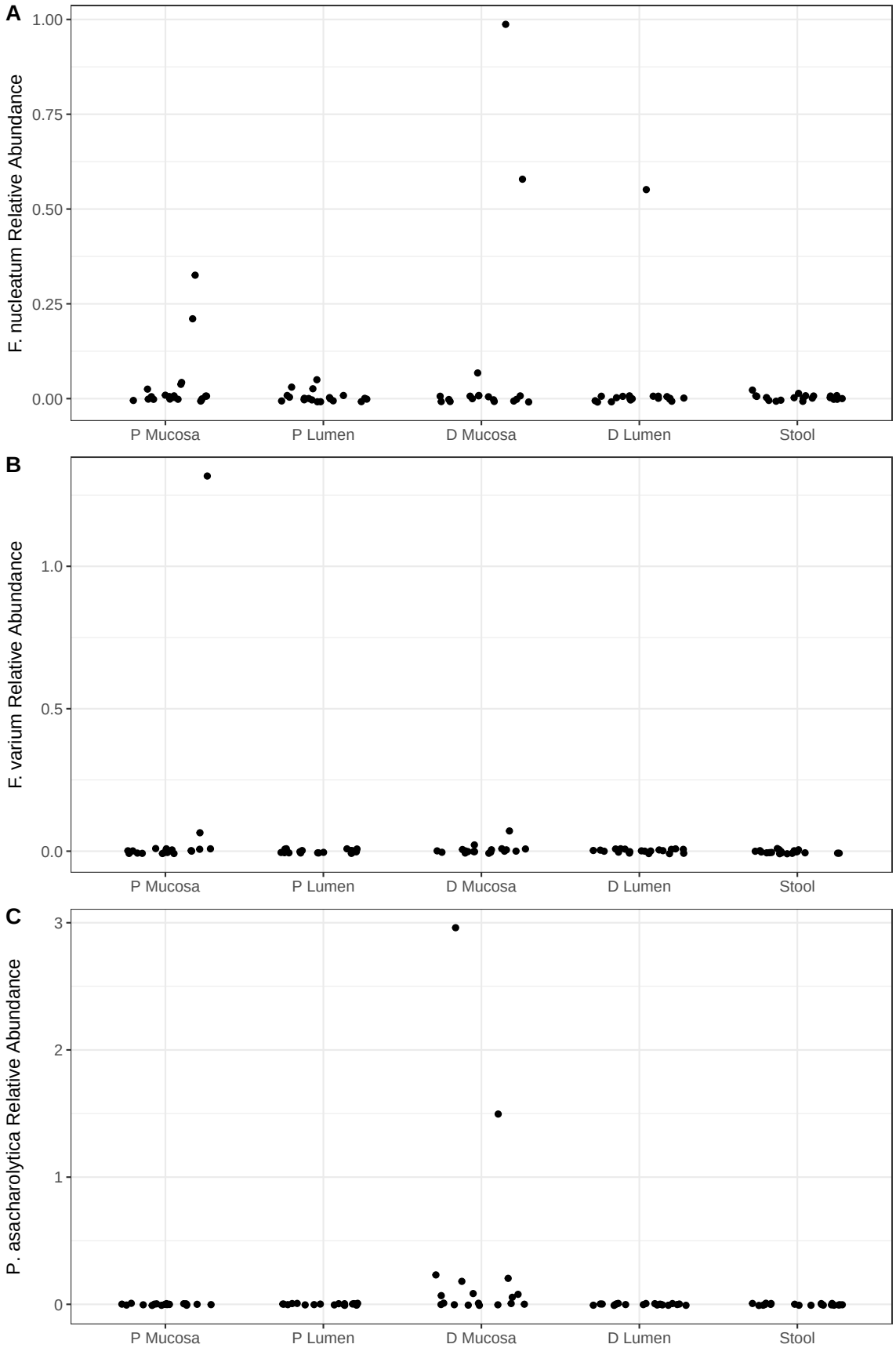
341 Taxa specific to the distal and proximal sides of the colon. Top five OTUs that are most important for
342 the classification model for the distal mucosa and lumen (A) and the proximal mucosa and lumen (B).



343

344 **Figure 5**

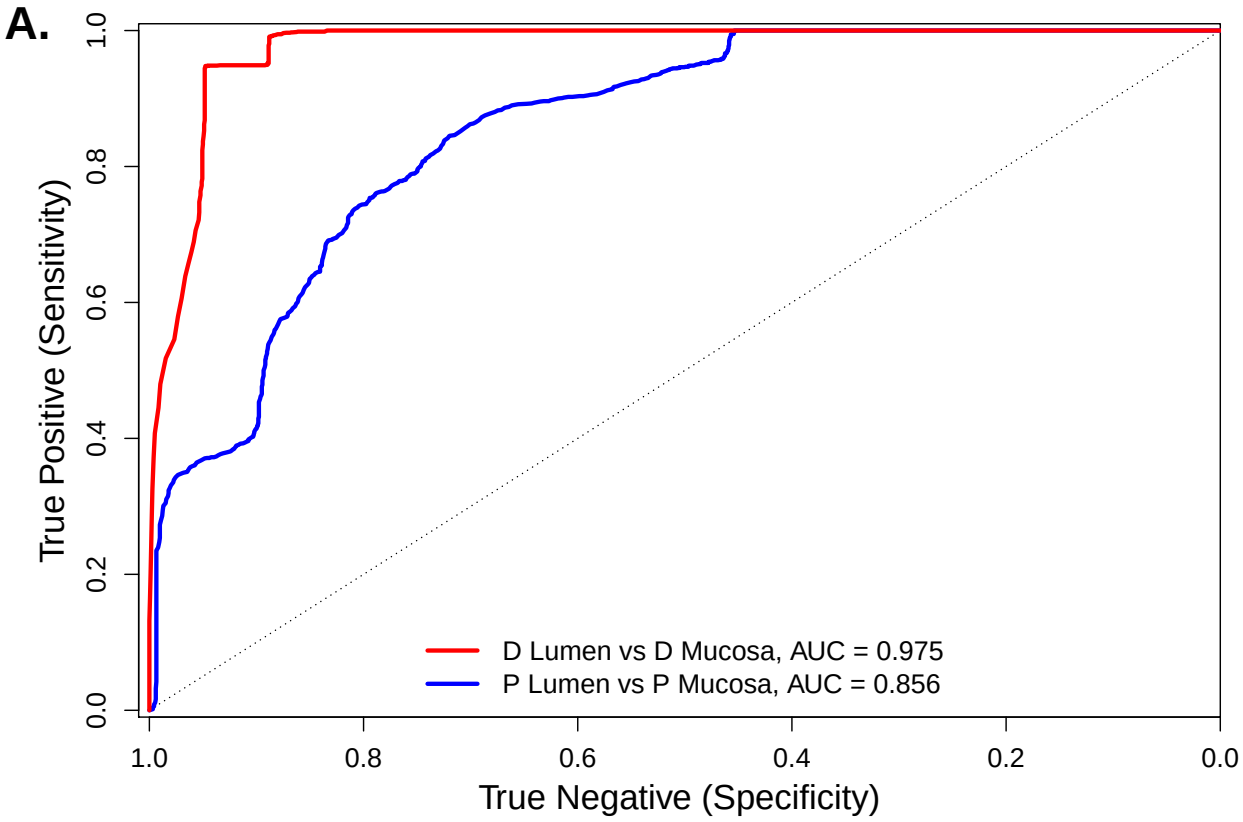
345 Taxa specific to the distal and proximal mucosa and lumen. Top five OTUs that are most important for
346 the classification model for the distal and proximal mucosa (A) and the distal and proximal lumen (B).

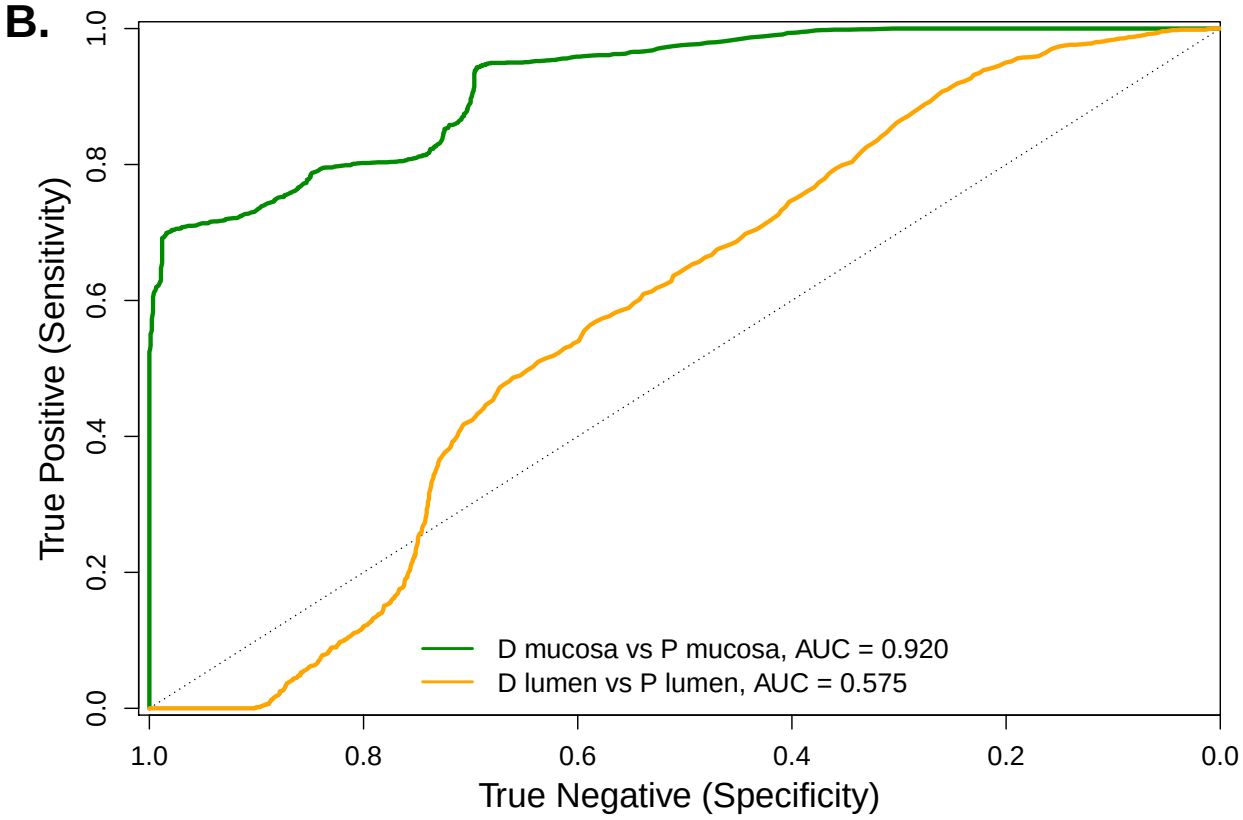

A
F. nucleatum Relative Abundance
1.00
0.75
0.50
0.25
0.00
P Mucosa
P Lumen
D Mucosa
D Lumen
Stool
B
F. varium Relative Abundance
1.0
0.5
0.0
P Mucosa
P Lumen
D Mucosa
D Lumen
Stool
C
P. asacharolytica Relative Abundance
3
2
1
0
P Mucosa
P Lumen
D Mucosa
D Lumen
Stool

347

Figure 6

Location and abundance of cancer-associated OTUs. Relative abundance was calculated and plotted by sample site for each OTU of interest: (A) *Fusobacterium nucleatum subsp. animalis* (B) *Fusobacterium varium* and (C) *Porphyromonas asacharolytica*


A.
True Positive (Sensitivity)
1.0
0.8
0.6
0.4
0.2
0.0
D Lumen vs D Mucosa, AUC = 0.975
P Lumen vs P Mucosa, AUC = 0.856
1.0
0.8
0.6
0.4
0.2
0.0
True Negative (Specificity)


B.
True Positive (Sensitivity)
1.0
0.8
0.6
0.4
0.2
0.0
D mucosa vs P mucosa, AUC = 0.920
D lumen vs P lumen, AUC = 0.575
1.0
0.8
0.6
0.4
0.2
0.0
True Negative (Specificity)

352

Figure S1

Random Forest classifies locations in the colon. A) Receiver Operator Characteristic curves are shown for the 10-fold cross validation of the Random Forest model classifying distal lumen versus proximal lumen (orange) and distal mucosa vs proximal mucosa (green). (B) Receiver Operator Characteristic curves are shown for the 10-fold cross validation of the Random Forest model classifying lumen and mucosal samples for the distal (red) and proximal (blue) sides of the colon.

References

1. **Yamauchi M**, **Lochhead P**, **Morikawa T**, **Huttenhower C**, **Chan AT**, **Giovannucci E**, **Fuchs C**, **Ogino S**. 2012. Colorectal cancer: A tale of two sides or a continuum?: Figure 1. Gut **61**:794–797. doi:10.1136/gutjnl-2012-302014.

2. **Forbes JD**, **Domselaar GV**, **Bernstein CN**. 2016. The gut microbiota in immune-mediated inflammatory diseases. Frontiers in Microbiology **7**. doi:10.3389/fmicb.2016.01081.

3. **Halfvarson J**, **Brislawn CJ**, **Lamendella R**, **Vazquez-Baeza Y**, **Walters WA**, **Bramer LM**, **DAmato M**, **Bonfiglio F**, **McDonald D**, **Gonzalez A**, **McClure EE**, **Dunklebarger MF**, **Knight R**, **Jansson JK**. 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. Nature Microbiology **2**:17004. doi:10.1038/nmicrobiol.2017.4.

4. **Benedix F**, **Kube R**, **Meyer F**, **Schmidt U**, **Gastinger I**, **Lippert H**. 2010. Comparison of 17, 641 patients with right- and left-sided colon cancer: Differences in epidemiology, perioperative course, histology, and survival. Diseases of the Colon & Rectum **53**:57–64. doi:10.1007/dcr.0b013e3181c703a4.

5. **Albenberg L**, **Esipova TV**, **Judge CP**, **Bittinger K**, **Chen J**, **Laughlin A**, **Grunberg S**, **Baldassano RN**, **Lewis JD**, **Li H**, **Thom SR**, **Bushman FD**, **Vinogradov SA**, **Wu GD**. 2014. Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. Gastroenterology **147**:1055–1063.e8. doi:10.1053/j.gastro.2014.07.020.

6. **Donaldson GP**, **Lee SM**, **Mazmanian SK**. 2015. Gut biogeography of the bacterial microbiota. Nature Reviews Microbiology **14**:20–32. doi:10.1038/nrmicro3552.

7. **Dejea CM**, **Wick EC**, **Hechenbleikner EM**, **White JR**, **Welch JLM**, **Rossetti BJ**, **Peterson**

SN, **Snesrud EC**, **Borisy GG**, **Lazarev M**, **Stein E**, **Vadivelu J**, **Roslani AC**, **Malik AA**, **Wanyiri JW**, **Goh KL**, **Thevambiga I**, **Fu K**, **Wan F**, **Llosa N**, **Housseau F**, **Romans K**, **Wu X**, **McAllister FM**, **Wu S**, **Vogelstein B**, **Kinzler KW**, **Pardoll DM**, **Sears CL**. 2014. Microbiota organization is a distinct feature of proximal colorectal cancers. Proceedings of the National Academy of Sciences **111**:18321–18326. doi:10.1073/pnas.1406199111.

8. **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**. doi:10.1186/s13073-016-0290-3.

9. **Strauss J**, **Kaplan GG**, **Beck PL**, **Rioux K**, **Panaccione R**, **DeVinney R**, **Lynch T**, **Allen-Vercoe E**. 2011. Invasive potential of gut mucosa-derived fusobacterium nucleatum positively correlates with IBD status of the host. Inflammatory Bowel Diseases **17**:1971–1978. doi:10.1002/ibd.21606.

10. **Jalanka J**, **Salonen A**, **Salojärvi J**, **Ritari J**, **Immonen O**, **Marciani L**, **Gowland P**, **Hoad C**, **Garsed K**, **Lam C**, **Palva A**, **Spiller RC**, **Vos WM de**. 2014. Effects of bowel cleansing on the intestinal microbiota. Gut **64**:1562–1568. doi:10.1136/gutjnl-2014-307240.

11. **Harrell L**, **Wang Y**, **Antonopoulos D**, **Young V**, **Lichtenstein L**, **Huang Y**, **Hanauer S**, **Chang E**. 2012. Standard colonic lavage alters the natural state of mucosal-associated microbiota in the human colon. PLoS ONE **7**:e32545. doi:10.1371/journal.pone.0032545.

12. **Lloyd-Price J**, **Abu-Ali G**, **Huttenhower C**. 2016. The healthy human microbiome. Genome Medicine **8**. doi:10.1186/s13073-016-0307-y.

13. **Eckburg PB**. 2005. Diversity of the human intestinal microbial flora. Science **308**:1635–1638. doi:10.1126/science.1110591.

14. **Cárcer DA de**, **Cuív PÓ**, **Wang T**, **Kang S**, **Worthley D**, **Whitehall V**, **Gordon I**, **McSweeney C**, **Leggett B**, **Morrison M**. 2010. Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. The ISME Journal **5**:801–809. doi:10.1038/ismej.2010.177.

15. **Zhang Z**, **Geng J**, **Tang X**, **Fan H**, **Xu J**, **Wen X**, **Ma Z (Sam)**, **Shi P**. 2013. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. The ISME Journal

406 **8**:881–893. doi:10.1038/ismej.2013.185.

407 16. **Castellarin M**, **Warren RL**, **Freeman JD**, **Dreolini L**, **Krzywinski M**, **Strauss J**, **Barnes R**,
408 **Watson P**, **Allen-Vercoe E**, **Moore RA**, **Holt RA**. 2011. Fusobacterium nucleatum infection is
409 prevalent in human colorectal carcinoma. Genome Research **22**:299–306. doi:10.1101/gr.126516.111.

410 17. **Lee Y**, **Eun CS**, **Lee AR**, **Park CH**, **Han DS**. 2016. FusobacteriumIsolates recovered from
411 colonic biopsies of inflammatory bowel disease patients in korea. Annals of Laboratory Medicine **36**:387.
412 doi:10.3343/alm.2016.36.4.387.

413 18. **Hong P-Y**, **Croix JA**, **Greenberg E**, **Gaskins HR**, **Mackie RI**. 2011. Pyrosequencing-based
414 analysis of the mucosal microbiota in healthy individuals reveals ubiquitous bacterial groups and micro-
415 heterogeneity. PLoS ONE **6**:e25042. doi:10.1371/journal.pone.0025042.

416 19. **Stearns JC**, **Lynch MDJ**, **Senadheera DB**, **Tenenbaum HC**, **Goldberg MB**, **Cvitkovitch DG**,
417 **Croitoru K**, **Moreno-Hagelsieb G**, **Neufeld JD**. 2011. Bacterial biogeography of the human digestive
418 tract. Scientific Reports **1**. doi:10.1038/srep00170.

419 20. **Sears CL**, **Garrett WS**. 2014. Microbes, microbiota, and colon cancer. Cell Host & Microbe
420 **15**:317–328. doi:10.1016/j.chom.2014.02.007.

421 21. **Mima K**, **Cao Y**, **Chan AT**, **Qian ZR**, **Nowak JA**, **Masugi Y**, **Shi Y**, **Song M**, **Silva A da**,
422 **Gu M**, **Li W**, **Hamada T**, **Kosumi K**, **Hanyuda A**, **Liu L**, **Kostic AD**, **Giannakis M**, **Bullman S**,
423 **Brennan CA**, **Milner DA**, **Baba H**, **Garraway LA**, **Meyerhardt JA**, **Garrett WS**, **Huttenhower C**,
424 **Meyerson M**, **Giovannucci EL**, **Fuchs CS**, **Nishihara R**, **Ogino S**. 2016. Fusobacterium nucleatum
425 in colorectal carcinoma tissue according to tumor location. Clinical and Translational Gastroenterology
426 **7**:e200. doi:10.1038/ctg.2016.53.

427 22. **Whitmore SE**, **Lamont RJ**. 2014. Oral bacteria and cancer. PLoS Pathogens **10**:e1003933.
428 doi:10.1371/journal.ppat.1003933.

429 23. **Flynn KJ**, **Baxter NT**, **Schloss PD**. 2016. Metabolic and community synergy of oral bacteria in
430 colorectal cancer. mSphere **1**:e00102–16. doi:10.1128/msphere.00102-16.

431 24. **Dharmani P**, **Strauss J**, **Ambrose C**, **Allen-Vercoe E**, **Chadee K**. 2011. Fusobacterium nucleatum
432 infection of colonic cells stimulates MUC2 mucin and tumor necrosis factor alpha. Infection and Immunity

433 **79**:2597–2607. doi:10.1128/iai.05118-11.

434 25. **Ohkusa T**. 2003. Induction of experimental ulcerative colitis by fusobacterium varium isolated from
435 colonic mucosa of patients with ulcerative colitis. Gut **52**:79–83. doi:10.1136/gut.52.1.79.

436 26. **Ohkusa T**, **Sato N**, **Ogihara T**, **Morita K**, **Ogawa M**, **Okayasu I**. 2002. Fusobacterium varium
437 localized in the colonic mucosa of patients with ulcerative colitis stimulates species-specific antibody.
438 Journal of Gastroenterology and Hepatology **17**:849–853. doi:10.1046/j.1440-1746.2002.02834.x.

439 27. **Shobar RM**, **Velineni S**, **Keshavarzian A**, **Swanson G**, **DeMeo MT**, **Melson JE**, **Losurdo J**,
440 **Engen PA**, **Sun Y**, **Koenig L**, **Mutlu EA**. 2016. The effects of bowel preparation on microbiota-related
441 metrics differ in health and in inflammatory bowel disease and for the mucosal and luminal microbiota
442 compartments. Clinical and Translational Gastroenterology **7**:e143. doi:10.1038/ctg.2015.54.

443 28. **Kozich JJ**, **Westcott SL**, **Baxter NT**, **Highlander SK**, **Schloss PD**. 2013. Development
444 of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on
445 the MiSeq illumina sequencing platform. Applied and Environmental Microbiology **79**:5112–5120.
446 doi:10.1128/aem.01043-13.

447 29. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**,
448 **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Horn DJV**, **Weber**
449 **CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for
450 describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541.
451 doi:10.1128/aem.01541-09.

452 30. **Wang Q**, **Garrity GM**, **Tiedje JM**, **Cole JR**. 2007. Naive bayesian classifier for rapid assignment of
453 rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**:5261–5267.
454 doi:10.1128/aem.00062-07.

455 31. **Liaw A**, **Wiener M**. 2002. Classification and regression by randomForest. R News: The Newsletter
456 of the R Project **2**:18–22.