Revisiting the Relationship between Short-Chain Fatty Acids, the Microbiota, and Colorectal Tumors

Running title: SCFAs and colorectal tumors

Marc A. Sze[1], Nicholas A. Lesniak[1], Mack T. Ruffin IV[2], Patrick D. Schloss[1†]

† To whom correspondence should be addressed: pschloss@umich.edu

1 Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

2 Department of Family Medicine and Community Medicine, Penn State Hershey Medical Center, Hershey, PA

Abstract

Background. Colorectal cancer (CRC) is increasing in prevalence in individuals under 50 and because of this will be a continuing health concern for the foreseeable future. The majority of the risk for developing CRC is attributable to environmental factors. One of these environmental factors is the microbiota, with certain bacterial community members being associated with CRC and other taxa being associated to individuals without tumors. Some of the bacterial species in taxa associated to individuals without tumors can use fiber to produce short-chain fatty acids (SCFAs) that inhibit tumor growth in model systems. However, the data supporting the importance of SCFAs in human CRC is less certain. Here, we test the hypothesis that SCFA concentrations and taxa associated with their production are different in individuals with colorectal tumors.

Methods. We analyzed a cross-sectional (n=490) and longitudinal pre- and post-treatment (n=67) group for their fecal concentrations of acetate, butyrate, and propionate. Analysis also included imputed gene relative abundance with PICRUSt, metagenomic sequencing on a subset (n=85) of the total cross-sectional group, and tumor classification and SCFA prediction models using Random Forest.

Results. No difference in SCFA concentrations were found between individuals without tumors and patients with adenomas or carcinomas (P-value > 0.15). Using metagenomic sequencing, there was also no difference in genes involved with SCFA synthesis between individuals without tumors and patients with adenomas or carcinomas (P-value > 0.70). Finally, there was no difference between the ability of Random Forest models to predict patients with adenomas or carcinomas versus individuals without tumors (P-value > 0.05).

Conclusions. Although our data does not support the hypothesis that fecal SCFA concentrations in patients in the general CRC population are different, there still may be specific types of colorectal tumors where SCFAs may be beneficial for treatment of CRC. Alternatively, our observations also support the hypothesis that there may be other metabolites or mechanisms (e.g. bacterial niche exclusion) that may be more protective against tumorigenesis and have not been thoroughly investigated in the context of human CRC.

Introduction

Colorectal cancer (CRC) is currently the third leading cancer-related cause of death within the US and the prevelance is increasing in invidiauls under 50 years of age (1, 2). Although there is a genetic component to the disease, the environment is considered a larger risk factor for CRC (3). These environmental risk factors include but are not limited to smoking cigarettes, diet, and the microbiota (4–6). Many of these environmental risk factors, including the microbiota, are modifiable. This has lead to the investigation of how the microbiota may exacerbate or cause tumorigensis (7–9) and whether the bacterial community is altered in CRC (10, 11). Many of these previous case/control studies have identified resident bacterial taxa to be decreased in patients with carcinoma tumors (11–13). Many of the bacterial species from these resident taxa identified in these case/control studies actively produce short-chain fatty acids (SCFAs) from fiber that are part of our general diet (14). The most extensively studied SCFAs are acetate, butyrate, and propionate (15). These SCFAs are hypothesized to be the main metabolites involved with protection against tumorigenesis and could help to reduce the risk of CRC.

Prior research suggests that SCFAs have promise in acting as an anti-tumorigenic agent. Specific SCFAs have shown positive results within model systems (16). For example, butyrate has been shown to inhibit cancer cell growth in *in vitro* systems (17). Additionally, fiber supplementation in mouse models of CRC caused an overall reduction in tumor burden while also increasing SCFA concentrations (18). These exciting results in model systems suggest that supplementation with food sources that bacteria use to create these SCFAs may confer beneficial effects against CRC. However, it is important to note that these model systems provide only preliminary evidence towards the ability of SCFAs to reduce and treat tumors and the studies reporting benefit in humans has been less convincing.

Overall, there is a lack of evidence on the benefit of increasing SCFA concentrations to protect against CRC in human populations. The initial case/control studies that investigated SCFA concentrations in CRC found that patients with carcinomas had lower concentrations of acetate, butyrate, and propionate versus patients with adenomas or individuals without colon tumors (19). Although this would argue that increasing SCFA concentrations could be protective against

tumorigenesis, fiber supplementation in randomized controlled trials have consistently failed to protect against tumor recurrence (20). These findings argue against the utility of treatments that aim to use SCFAs to reduce or protect against tumorigenesis. Given the lack of clear evidence in human studies of the benefit of SCFAs in CRC, there is a need for more investigation into this area.

Our study fills some of the current gaps in the literature that relate to the study of SCFAs and CRC in human populations. Specifically, by using a seperate cohort it tests previous case/control findings on SCFA concentrations in individuals with and without tumors. Additionally, prior investigations grouped patients with adenoma and individuals without tumors into a single group to compare against patients with carcinomas. Despite doing this, the suggestion was made that a reduction in SCFA concentration would also be observed between patients with adenomas and indivdiuals without tumors. We also can test this suggestion because we have a larger number of patients with adenomas within our study and do not need to group patients with adenomas and individuals without tumors into a single group. Additionally, we build upon these observations by assessing the utility of using SCFAs and Operational Taxonomic Units (OTUs) as a risk stratification tool of colorectal tumors (adenoma or carcinoma). We also investigate whether OTUs that are most important to these models are closely associated with the classification of SCFA concentrations. Collectively, this study provides important information on the replicability of previous findings in humans by extensively studying how SCFAs are associated with colorectal tumors.

To accomplish this task we directly measured the concentration of acetate, butyrate, and propionate within fecal samples for two different groups. The first group had a sample obtained at a single cross sectional point in time while the second group had samples obtained before (pre-) and after (post-)treatment for colorectal tumors. To provide further support for our SCFA findings we also used PICRUSt (21) and metagenomic sequencing to investigate differences in genes involved with SCFA synthesis between individuals without tumors, patients with adenomas, and patients with carcinomas. Next, we investigated whether taxa associated with SCFA production were important to disease classification models. First, using the cross-sectional data, we analyzed the number of correlations between OTU relative abundance and SCFA concentrations across individuals without tumors and patients with adenomas or carcinomas. Second, we assessed the affect adding SCFA concentrations to OTU data had on classification of patients with adenomas or carcinomas using the

Random Forest algorithm (22). Third, we analyzed how well 16S rRNA gene sequencing predicts SCFA concentrations. Finally, we compared whether SCFA concentrations replace important taxa in disease models and whether the same OTUs are the most important variables used to classify disease and SCFA concentration. Collectively, this investigation provides additional information as to whether SCFAs are decreased in patients with colorectal tumors and provides context as to whether targeting taxa to increase SCFA concentrations is a viable option to protect against colon tumorigenesis.

Results

Decreased SCFA concentrations are not associated with adenoma or carcinoma tumors. We used high-performance liquid chromatography (HPLC) to measure acetate, butyrate, and propionate concentrations of frozen fecal samples from 490 individuals at a cross-sectional point in time. There was no difference between individuals without colon tumors (n=172) and patients with either an adenoma (n=198) or carcinoma (n=120) for any of the SCFAs measured after multiple comparison correction (P-value > 0.15) [Figure 1A - 1C]. We next measured the concentration of SCFAs in 67 patients with an adenoma (n=41) or carcinoma (n=26) in which we had pre- and post-treatment fecal samples. Although there was a general trend for increasing acetate, butyrate, and propionate concentrations following treatment for tumors, there was no significant difference between pre- and post-treatment for either patients with adenomas (P-value > 0.20) or carcinomas (P-value > 0.80) [Figure 1D - 1F].

Gene abundance for enzymes involved in SCFA synthesis are the same for individuals without tumors and patients with adenomas or carcinomas. In order to provide further confirmation and support of our SCFA concentration results we investigated the genes encoding specific enzymes involved with SCFA synthesis [Table S1]. Using this list of specific genes [Table S1], we looked for differences in gene abundance between individuals without colon tumors and patients with adenomas or carcinomas. Although we intended to analyze all the genes in the list [Table S1], not all of the KEGG genes were identified during our analysis. We first analyzed imputed gene relative abundance, calculated from our OTU data that was generated from 16S rRNA gene sequencing. We found no difference in any of the imputed gene relative abundances for enzymes involved with acetate, butyrate, or propionate synthesis (P-value > 0.90) [Table S2]. Since butyrate is one of the most studied SCFAs in the context of CRC, we visualized the observed lack of difference between indviduals without tumors and patients with adenomas or carcinomas in gene abundance for enzymes involved with SCFA synthesis using the butyrate kinase gene [Figure 2A]. Additionally, using a paired Wilcoxon rank-sum test, there also was no difference in imputed gene relative abundance between pre- and post-treatment samples for any genes involved with SCFA synthesis (P-value > 0.70) [Table S3]. Next, we took a subset of these 490 fecal samples

(n=85) and used metagenomic sequencing to confirm these results. Like the imputed gene results, metagenomic analysis found that there was no difference in any of the genes involved in SCFA synthesis between individuals without colon tumors (n=29) and patients with adenomas (n=28) or carcinomas (n=28) (P-value > 0.70) [Table S4]. This similarity between indviduals without tumors and patients with adenomas or carcinomas is highlighted again by visualizing the gene abundance for the butyrate kinase gene [Figure 2B]. These observations provide evidence that gene prevalence for enzymes involved in SCFA synthesis does not change due to colorectal tumors and provides further support for our original SCFA concentration observations.

The number of OTUs positively associated with SCFA concentration were similar between individuals without tumors and patients with adenomas or carcinomas. Having found no difference between individuals without tumors and patients with adenomas or carcinomas in SCFA concentrations or genes encoding enzymes involved with SCFA synthesis, we next investigated if specific OTUs correlated with SCFA concentrations. The main goal of this analysis was to identify if there were OTUs that were significantly associated with SCFA concentrations and if this was different between individuals that did not have tumors, had an adenoma, or had a carcinoma. To accomplish this we used Spearmans rho, a non-parametric measure of association, and tested if there was a correlation that was signficantly greater than zero. We found that taxa from *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* dominated statistically significant OTU correlations [Figure 3 & Table S5]. There was a noticeably higher number of significant negative correlations associated with patients with adenomas for all SCFAs tested [Figure 3]. In particular, OTUs from the *Ruminococcaceae* family had the largest share of these negative correlations within patients with adenomas [Figure 3]. Patients with adenomas also had more positive correlations between OTUs and SCFA concentrations, but their total number was more similar to individuals without tumors or patients with carcinomas versus the analogous comparison for the number of negative correlations [Figure 3]. Additionally, the number of positive correlations between OTUs and SCFA concentrations was similar between individuals without tumors and patients with a carcinoma [Figure 3]. Finally, when we used high/low SCFA groups based on the overall median concentration for each SCFA instead of SCFA concentrations a similar pattern was still observed [Figure S1 & Table S6]. Overall, these results suggest that the resident taxa that may change the most due

to colon tumors may not be ones that are responsible for the production of acetate, butyrate, or propionate.

SCFA concentrations do not replace important *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* OTUs in Random Forest models built to classify tumors. Despite the lack of difference in positive correlations between OTUs and SCFA concentrations between individuals with and without tumors, OTUs associated with SCFA concentrations could still be the most important variables to Random Forest models built to classify patients with adenomas or carcinomas. We tested this by using the Random Forest algorithm to build models with OTU abundance data or OTU abundances and SCFA concentrations to classify normal versus adenoma and normal versus carcinoma fecal samples. With these models we compared whether OTUs with taxonomic classification to *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* remained when SCFA concentrations were added to the model. Additionally, we also compared whether any of the important OTUs that remained also had significant correlations with SCFA concentrations. Both our adenoma and carcinoma models classified patients with a similar degree of success, as measured by the area under the curve (AUC) (P-value > 0.05) [Figure 4A & 4D]. After the addition of SCFA concentrations to the adenoma or carcinoma models, many OTUs with taxonomic classification to *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* remained as important variables to the model [Figure Figure 4B-C & 4E-F]. After adding SCFA concentrations to the adenoma model, there were only 2 OTUs significantly associated with SCFA concentration that were part of the top 10 most important variables and both were postively associated with acetate, butyrate, or propionate concentrations. When SCFA concentrations were added to the carcinoma model, only 1 OTU that was associated with SCFA concentration remained as part of the top 10 most important variables and it was negatively associated with acetate and butyrate concentrations. In combination with the previous results on OTU correlations, these observations provide additional evidence that the resident taxa that are associated with protection against tumorigeneis are not ones associated with acetate, butyrate, or propionate production.

The most important OTUs in Random Forest models built to classify SCFA concentrations or tumors are not the same. It is possible that due to the way the Random Forest algorithm works, OTUs associated with SCFA concentrations could be downweighted in importance within

the adenoma or carcinoma models when SCFA concentrations are included. To test if this is the case we used OTU data and built Random Forest models to classify SCFA concentrations. Overall, the correlation between the predicted and actual SCFA concentrations were moderately associated with each other [Figure 5A]. Additionally, because the training set R^2 was always higher than the test set R^2, all SCFA concentration models tended to be over fit, suggesting that rarer taxa were important for these classifications [Figure 5A]. There also was a difference in accuracy based on whether the fecal sample was from an individual without tumors or from patients with adenomas or carcinomas [Figure 5B]. When comparing the adenoma or carcinoma model to the SCFA concentration models there was minimal overlap between these model's most important OTUs [Figure 4B-C, 4E-F and 5C-E]. The only OTU that was in the top 10 most important variables and had overlap between the models was OTU00167 (*Clostridiales*) [Figure 4B-C, 4E-F, 5C-E]. Similar observations were made when using high/low SCFA groups based on the median SCFA concentration [Figure S2]. Collectively, these observations provide evidence that it is possible to identify specific OTUs associated with higher SCFA concentrations and accordingly these OTUs belong to taxa known to produce acetate, butyrate, and propionate. Although it is possible to identify OTUs associated with SCFA production, overall, our results do not support the hypothesis that SCFA concentration or OTUs associated with their production are different between individuals with no tumors and patients with adenomas or carcinomas.

Discussion

The observations from this study do not support the hypothesis that SCFA concentrations are different in individuals with tumors. Whether we directly measured the SCFA concentration or investigated genes that encoded enzymes used for their production, no difference could be identified between individuals without tumors and patients with adenomas or carcinomas [Figure 1 & 2]. Although there were differences in the number of significant correlations between SCFA concentration and OTU relative abundance based on whether individuals did not have tumors, had an adenoma, or had a carcinoma, SCFA concentrations did not provide increased model accuracy for tumor classification [Figure 3-4 & S1]. In models with SCFA concentrations included, many OTUs that classified to *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* remained as important variables for the model [Figure 4]. Additionally, when models using OTU relative abundance to classify SCFA concentrations were assessed, the OTUs that classified to *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* were not the same as the OTUs that classified to these taxa in the tumor models [Figure 4-5]. Collectively, our observations suggest that the resident taxa from *Clostridiales*, *Lachnospiraceae*, and *Ruminococcaceae* that are different between individuals without tumors and patients with adenomas or carcinomas, are not the same as those involved with SCFA production.

Although SCFAs have been shown to be anti-tumorigenic, most of these studies have been performed in model systems (16, 17). Many of the *in vivo* studies use proxies such as fiber supplementation rather than SCFAs directly (14). Although it is well known that breakdown products from gut bacteria results in SCFA production, fiber effects on tumorigenesis may be through other mechanisms in these *in vivo* model systems. Additionally, the observations in humans on the benefit of SCFAs in preventing tumorigenesis have been mixed. In previous case/control studies lower SCFA concentrations were observed in patients with carcinomas versus those without carcinomas (19). Yet, this is in contrast to multiple randomized-controlled trials that have found no difference in tumor recurrence between patients who do and do not get fiber supplementation (20, 23). In contrast to the *in vivo* model findings, the observations made in these randomized-controlled trials would suggest that SCFAs do not prevent or slow tumorigenesis. One reason for these results is

that SCFA concentrations and responses to fiber vary quite a bit between healthy individuals (24). This information taken together with our observations would suggest that either individuals who do not respond to fiber supplementation would need to acquire these bacteria to achieve a benefit or that SCFAs provide little to no benefit as an anti-tumorigenic compound in colorectal cancer.

Another possible alternative explanation as to why no difference in SCFA concentration between individuals without tumors and patients with adenomas or carcinomas was observed, could be because only certain types of colorectal cancers are affected by SCFAs. One limitation of current research into the effect of SCFAs and the microbiota in CRC has been that all tumors are treated as the same type. However, there are known differences in the types of mutations that occur (25) and treating all tumors as equal may actually hide any benefit that could be found in certain subsets of individuals. Similar to the idea of using immunotherapy as a targeted treatment option for specific tumors (26), SCFAs may have beneficial effects for distinct types of colorectal tumors. Future research will need to test if this is a valid hypothesis. Regardless of this limitation, our results in combination to previous randomized controlled trials on fiber supplementation (20, 23) suggests that using SCFAs as a general treatment for colorectal cancer is unlikely to provide a reduction in tumorigenesis.

One possible technical limitation is that a fecal sample may not be an ideal type of bio-specimen and that the effect SCFAs have on tumorigenesis is only detected in the colon. However, this is unlikely to be a major confounder. First, most *in vivo* studies as well as human studies have used fecal material in their analysis (18, 19). Second, previous studies that measure SCFA changes after fiber supplementation use fecal material to track these responses with a great deal of success (24). Although there are limitations with the current research on SCFAs and colorectal tumors, technical limitations are less likely to be cause of this. Additionally, as mentioned earlier, our observations along with the randomized controlled trials on fiber supplementation in tumor recurrence (20) provide evidence that these specific metabolites may not be protective or used as a general treatment option in colorectal cancer. Yet, taxa that are associated with SCFA production are consistently higher in individuals without colon tumors than patients with carcinomas (10, 11, 27).

The potential protection against colorectal cancers may not be from SCFAs, even though taxa

associated with their production are higher in individuals without tumors versus patients with carcinomas (10, 11, 27). Indeed our data would support the contention that the taxa are similar to those associated with SCFA production but that these specific microbes or OTUs themselves are not associated with SCFAs. In particular, our results showing that different OTUs from the same taxonomic classification are in tumor and SCFA Random Forest models supports this hypothesis. This leads to the possibility that protection may be through two other routes. First, there could be a different pathway or other less extensively studied metabolites that provides the necessary protection against tumorigenesis. Alternatively, protection may not occur via a metabolite but instead through niche exclusion of mouth-associated microbes (e.g. *Fusobacterium*, *Porphyromonas*, *Parvimonas*, *Peptostreptococcus* (6, 12, 13)). The idea of niche exclusion is similar to how the community protects against *Clostridium difficile* infection (28) with chronic inflammation replacing the role of antibiotics. Even though we did not find lower concentrations of SCFAs associated with colorectal tumors, we think that there are many exciting new avenues to explore because of these results.

Conclusions

Our observations found no difference in SCFA concentration, their utility as a classification tool, or for genes of enzymes involved in SCFA synthesis between individuals without colon tumors and patients with either adenoma or carcinoma tumors. Although these results are different than other reports in the literature, they do align with the randomized controlled trials that have tested fiber use in preventing colorectal tumor recurrence. Additionally, these observations suggest that resident microbes that are not involved in SCFA production may be the important resident community members involved with preventing tumorigenesis. By focusing on the alternative mechanisms that are associated with these non-SCFA producing resident microbes, the identification of more promising therapeutic options for use in treating colorectal cancer may be found.

Materials and Methods

Study design and sampling. The overall protocol has been described in detail previously (29, 30). In brief, this study used fecal samples obtained at either a single cross-sectional time point (n=490) or from before (pre-) and after (post-) treatment of a patient's tumor (adenoma n =41 and carcinoma n = 26). For patients undergoing treatment for their tumor the length of time between their initial and follow up sample ranged from 188 - 546 days. Our use of treatment has been previously defined as encompassing removal of a tumor with or without chemotherapy and radiation (29). Diagnosis of tumor was made by colonoscopic examination and histopathological review of biopsies obtained (29, 30). The University of Michigan Institutional Review Board approved the study and informed consent was obtained from all participants in accordance to the guidelines set out by the Helsinki Declaration.

Measuring specific SCFAs. Our protocol for the measurement of acetate, butyrate, and propionate followed a previously published protocol that used a High-Performance Liquid Chromatography (HPLC) machine (24). The following changes to this protocol included the use of frozen fecal samples suspended in 1ml of PBS instead of fecal suspensions in DNA Genotek OmniGut tubes, and the use of the actual weight of fecal samples instead of the average weight for SCFA concentration normalizations. These methodological changes did not affect the overall median concentrations of these SCFAs between the two studies (see Table 1 (24) and Figure 1 here).

16s rRNA gene sequencing. The workflow and processing have been previously described (29, 31, 32). In brief, sequences were quality filtered and contigs created from the paired end reads. Any sequences with ambiguous base calls were discarded. Contigs were then checked for matches to the V4 region of the 16S rRNA gene using the SILVA database (33). Chimeras were identified and removed using UCHIME and OTUs clustered at 97% similarity (34). The major differences from these previous reports include: the use of version 1.39.5 of the mothur software package and clustering Operational Taxonomic Units (OTUs) at 97% similarity using the OptClust algorithm (35).

Generating imputed metagenomes. The use of PICRUSt version 1.1.2 with the recommended standard operating protocol (21) was used. Briefly, the mothur shared file and metadata was

converted into a biom formatted table using the biom convert function, the subsequent biom file was processed with the 'normalize_by_copy_number.py' function, and subsequent imputed metagenomes created using the 'predict_metagenomes.py' function.

Obtaining Operational Protein Families from metagenomes. A subset of the cross-sectional group (n=490) containing a total of 85 individuals (normal n=29, adenoma n=28, and carcinoma n=28) was shotgun sequenced on an Illumina HiSeq using 125 bp paired end reads and a previously described method (36). Briefly, the sequences were quality filtered and sequences aligning to the human genome were removed prior to contig assembly with MEGAHIT (37). Open Reading Frames (ORFs) were identified using Prodigal (38), counts generated using Diamond (39), subsequent clustering into Operational Protein Families (OPFs) used mmseq2 (40), and OPF alignment used the KEGG database (41).

Pulling genes involved with SCFA synthesis. Specific genes located near the end of the pathways involved in the synthesis of acetate, butyrate, and propionate were analyzed for any differences between individuals with normal colons and those with tumors. These genes were based on pathways from KEGG as well as previous research (41, 42) and a list can be found in the supplemental material [Table S1].

Random Forest models. The model was first trained on 80% of the data and then tested on the held out 20% (80/20 split) using the Random Forest algorithm for classification and regression models via the caret package (22, 43). This was repeated on 100 different 80/20 splits of the data to generate a reasonable range for the AUC of the model. The reported AUCs, unless otherwise specified, are for the test sets. The classification models were built to group normal versus adenoma, normal versus carcinoma, and high versus low SCFA concentrations. The regression models were built to classify the SCFA concentrations of acetate, butyrate, and propionate regardless of disease status.

Statistical analysis workflow. All analysis was performed using the statistical language R (44). Generally, a Kruskal-Walis rank sum test with a Dunn's post-hoc test was used to assess differences between the groups used. Where appropriate Benjamini-Hochberg was used to correct for multiple comparisons (45). First, we assessed differences in SCFA concentrations measured by HPLC

between individuals with normal colons and patients with tumors (adenoma or carcinoma). We then analyzed whether SCFA concentrations changed in patients with an adenoma or carcinoma pre- versus post-treatment. Next, the imputed gene counts of important mediators of SCFA synthesis was tested. Additionally, the counts generated for OPFs that matched important genes involved with SCFA creation were analyzed. From here we analyzed the number of significant positive and negative correlations between OTU relative abundance and SCFA concentrations in individuals without tumors and patients with adenomas or carcinomas using Spearman's rho. Next, we assessed whether OTUs alone or OTUs and SCFAs were better able to classify individuals with and without tumors using Random Forest models. Finally, models to classify high or low SCFA concentration based on the median of each SCFA or the actual concentration using 16S rRNA gene sequencing data was created using the Random Forest algorithm. For all Random Forest models, the assessment of the most important variables was based on the top 10 features (OTUs or SCFAs) using the mean decrease in accuracy.

Acknowledgements

The authors thank the Great Lakes-New England Early Detection Research Network for providing the fecal samples that were used in this study. We would also like to thank Kwi Kim and Thomas M Schmidt for their help in running the short-chain fatty acid analysis on the High-Performance Liquid Chromatography machine at the University of Michigan. Salary support for Marc A. Sze came from the Canadian Institute of Health Research and NIH grant UL1TR002240. Salary support for Patrick D. Schloss came from NIH grants P30DK034933 and 1R01CA215574.

References

1. **Haggar F**, **Boushey R**. 2009. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. Clinics in Colon and Rectal Surgery **22**:191–197. doi:10.1055/s-0029-1242458.

2. **Siegel RL**, **Miller KD**, **Jemal A**. 2016. Cancer statistics, 2016. CA: A Cancer Journal for Clinicians **66**:7–30. doi:10.3322/caac.21332.

3. **Lichtenstein P**, **Holm NV**, **Verkasalo PK**, **Iliadou A**, **Kaprio J**, **Koskenvuo M**, **Pukkala E**, **Skytthe A**, **Hemminki K**. 2000. Environmental and heritable factors in the causation of cancer analyses of cohorts of twins from sweden, denmark, and finland. New England Journal of Medicine **343**:78–85. doi:10.1056/nejm200007133430201.

4. **Fliss-Isakov N**, **Zelber-Sagi S**, **Webb M**, **Halpern Z**, **Kariv R**. 2017. Smoking habits are strongly associated with colorectal polyps in a population-based case-control study. Journal of Clinical Gastroenterology 1. doi:10.1097/mcg.0000000000000935.

5. **Lee J**, **Jeon JY**, **Meyerhardt JA**. 2015. Diet and lifestyle in survivors of colorectal cancer. Hematology/Oncology Clinics of North America **29**:1–27. doi:10.1016/j.hoc.2014.09.005.

6. **Kostic AD**, **Gevers D**, **Pedamallu CS**, **Michaud M**, **Duke F**, **Earl AM**, **Ojesina AI**, **Jung J**, **Bass AJ**, **Tabernero J**, **Baselga J**, **Liu C**, **Shivdasani RA**, **Ogino S**, **Birren BW**, **Huttenhower C**, **Garrett WS**, **Meyerson M**. 2011. Genomic analysis identifies association of fusobacterium with colorectal carcinoma. Genome Research **22**:292–298. doi:10.1101/gr.126573.111.

7. **Zackular JP**, **Baxter NT**, **Iverson KD**, **Sadler WD**, **Petrosino JF**, **Chen GY**, **Schloss PD**. 2013. The gut microbiome modulates colon tumorigenesis. mBio **4**:e00692–13–e00692–13. doi:10.1128/mbio.00692-13.

8. **Baxter NT**, **Zackular JP**, **Chen GY**, **Schloss PD**. 2014. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. Microbiome **2**:20. doi:10.1186/2049-2618-2-20.

9. **Zackular JP**, **Baxter NT**, **Chen GY**, **Schloss PD**. 2015. Manipulation of the gut microbiota

reveals role in colon tumorigenesis. mSphere **1**:e00001–15. doi:10.1128/msphere.00001-15.

10. **Shah MS**, **DeSantis TZ**, **Weinmaier T**, **McMurdie PJ**, **Cope JL**, **Altrichter A**, **Yamal J-M**, **Hollister EB**. 2017. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. Gut **67**:882–891. doi:10.1136/gutjnl-2016-313189.

11. **Sze MA**, **Schloss PD**. 2018. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. doi:10.1101/285486.

12. **Zeller G**, **Tap J**, **Voigt AY**, **Sunagawa S**, **Kultima JR**, **Costea PI**, **Amiot A**, **Bohm J**, **Brunetti F**, **Habermann N**, **Hercog R**, **Koch M**, **Luciani A**, **Mende DR**, **Schneider MA**, **Schrotz-King P**, **Tournigand C**, **Nhieu JTV**, **Yamada T**, **Zimmermann J**, **Benes V**, **Kloor M**, **Ulrich CM**, **Knebel Doeberitz M von**, **Sobhani I**, **Bork P**. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology **10**:766–766. doi:10.15252/msb.20145645.

13. **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.

14. **Holscher HD**. 2017. Dietary fiber and prebiotics and the gastrointestinal microbiota. Gut Microbes **8**:172–184. doi:10.1080/19490976.2017.1290756.

15. **Louis P**, **Flint HJ**. 2016. Formation of propionate and butyrate by the human colonic microbiota. Environmental Microbiology **19**:29–41. doi:10.1111/1462-2920.13589.

16. **O'Keefe SJD**. 2016. Diet, microorganisms and their metabolites and colon cancer. Nature Reviews Gastroenterology & Hepatology **13**:691–706. doi:10.1038/nrgastro.2016.165.

17. **Encarnação JC**, **Pires AS**, **Amaral RA**, **Gonçalves TJ**, **Laranjo M**, **Casalta-Lopes JE**, **Gonçalves AC**, **Sarmento-Ribeiro AB**, **Abrantes AM**, **Botelho MF**. 2018. Butyrate, a dietary fiber derivative that improves irinotecan effect in colon cancer cells. The Journal of Nutritional Biochemistry **56**:183–192. doi:10.1016/j.jnutbio.2018.02.018.

18. **Bishehsari F**, **Engen P**, **Preite N**, **Tuncil Y**, **Naqib A**, **Shaikh M**, **Rossi M**, **Wilber S**, **Green**

S, **Hamaker B**, **Khazaie K**, **Voigt R**, **Forsyth C**, **Keshavarzian A**. 2018. Dietary fiber treatment corrects the composition of gut microbiota, promotes SCFA production, and suppresses colon carcinogenesis. Genes **9**:102. doi:10.3390/genes9020102.

19. **Ohigashi S**, **Sudo K**, **Kobayashi D**, **Takahashi O**, **Takahashi T**, **Asahara T**, **Nomoto K**, **Onodera H**. 2013. Changes of the intestinal microbiota, short chain fatty acids, and fecal pH in patients with colorectal cancer. Digestive Diseases and Sciences **58**:1717–1726. doi:10.1007/s10620-012-2526-4.

20. **Yao Y**, **Suo T**, **Andersson R**, **Cao Y**, **Wang C**, **Lu J**, **Chui E**. 2017. Dietary fibre for the prevention of recurrent colorectal adenomas and carcinomas. Cochrane Database of Systematic Reviews. doi:10.1002/14651858.cd003430.pub2.

21. **Langille MGI**, **Zaneveld J**, **Caporaso JG**, **McDonald D**, **Knights D**, **Reyes JA**, **Clemente JC**, **Burkepile DE**, **Thurber RLV**, **Knight R**, **Beiko RG**, **Huttenhower C**. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology **31**:814–821. doi:10.1038/nbt.2676.

22. **Liaw A**, **Wiener M**. 2002. Classification and regression by randomForest. R News **2**:18–22.

23. **Schatzkin A**, **Lanza E**, **Corle D**, **Lance P**, **Iber F**, **Caan B**, **Shike M**, **Weissfeld J**, **Burt R**, **Cooper MR**, **Kikendall JW**, **Cahill J**, **Freedman L**, **Marshall J**, **Schoen RE**, **Slattery M**. 2000. Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas. New England Journal of Medicine **342**:1149–1155. doi:10.1056/nejm200004203421601.

24. **Venkataraman A**, **Sieber JR**, **Schmidt AW**, **Waldron C**, **Theis KR**, **Schmidt TM**. 2016. Variable responses of human microbiomes to dietary supplementation with resistant starch. Microbiome **4**. doi:10.1186/s40168-016-0178-x.

25. **Fearon ER**, **Vogelstein B**. 1990. A genetic model for colorectal tumorigenesis. Cell **61**:759–767. doi:10.1016/0092-8674(90)90186-i.

26. **Thomas X**, **Heiblig M**. 2016. The development of agents targeting the BCR-ABL tyrosine kinase as philadelphia chromosome-positive acute lymphoblastic leukemia treatment. Expert

Opinion on Drug Discovery **11**:1061–1070. doi:10.1080/17460441.2016.1227318.

27. **Sze MA**, **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2017. Normalization of the microbiota in patients after treatment for colonic lesions. Microbiome **5**. doi:10.1186/s40168-017-0366-3.

28. **Theriot CM**, **Young VB**. 2015. Interactions between the gastrointestinal microbiome and clostridium difficile. Annual Review of Microbiology **69**:445–461. doi:10.1146/annurev-micro-091014-104115.

29. **Sze MA**, **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2017. Normalization of the microbiota in patients after treatment for colonic lesions. Microbiome **5**. doi:10.1186/s40168-017-0366-3.

30. **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.

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

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

33. **Quast C**, **Pruesse E**, **Yilmaz P**, **Gerken J**, **Schweer T**, **Yarza P**, **Peplies J**, **Glöckner FO**. 2012. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research **41**:D590–D596. doi:10.1093/nar/gks1219.

34. **Edgar RC**, **Haas BJ**, **Clemente JC**, **Quince C**, **Knight R**. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27**:2194–2200. doi:10.1093/bioinformatics/btr381.

35. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an improved method for assigning

452 amplicon-based sequence data to operational taxonomic units. mSphere **2**:e00073–17.
453 doi:10.1128/mspheredirect.00073-17.

454 36. **Hannigan GD**, **Duhaime MB**, **Ruffin MT**, **Koumpouras CC**, **Schloss PD**. 2017. Diagnostic
455 potential & the interactive dynamics of the colorectal cancer virome. doi:10.1101/152868.

456 37. **Li D**, **Liu C-M**, **Luo R**, **Sadakane K**, **Lam T-W**. 2015. MEGAHIT: An ultra-fast single-node
457 solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics
458 **31**:1674–1676. doi:10.1093/bioinformatics/btv033.

459 38. **Hyatt D**, **Chen G-L**, **LoCascio PF**, **Land ML**, **Larimer FW**, **Hauser LJ**. 2010. Prodigal:
460 Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**:119.
461 doi:10.1186/1471-2105-11-119.

462 39. **Buchfink B**, **Xie C**, **Huson DH**. 2014. Fast and sensitive protein alignment using DIAMOND.
463 Nature Methods **12**:59–60. doi:10.1038/nmeth.3176.

464 40. **Steinegger M**, **Söding J**. 2017. MMseqs2 enables sensitive protein sequence searching for
465 the analysis of massive data sets. Nature Biotechnology. doi:10.1038/nbt.3988.

466 41. **Kanehisa M**, **Sato Y**, **Kawashima M**, **Furumichi M**, **Tanabe M**. 2015. KEGG as a
467 reference resource for gene and protein annotation. Nucleic Acids Research **44**:D457–D462.
468 doi:10.1093/nar/gkv1070.

469 42. **Baxter NT**, **Zackular JP**, **Chen GY**, **Schloss PD**. 2014. Structure of the gut microbiome
470 following colonization with human feces determines colonic tumor burden. Microbiome **2**:20.
471 doi:10.1186/2049-2618-2-20.

472 43. **Jed Wing MKC from**, **Weston S**, **Williams A**, **Keefer C**, **Engelhardt A**, **Cooper T**, **Mayer Z**,
473 **Kenkel B**, **R Core Team**, **Benesty M**, **Lescarbeau R**, **Ziem A**, **Scrucca L**, **Tang Y**, **Candan C**,
474 **Hunt. T**. 2017. Caret: Classification and regression training.

475 44. **R Core Team**. 2017. R: A language and environment for statistical computing. R Foundation

for Statistical Computing, Vienna, Austria.

45. **Benjamini Y**, **Hochberg Y**. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) **57**:289–300.

Figure 1. No change in SCFA measurements was observed between normal, adenoma, and carcinoma individuals using HPLC. Acetate concentrations in fecal samples of individuals without colon tumors, adenomas, and carcinomas (A). Butyrate concentrations in fecal samples of individuals without colon tumors, adenomas, and carcinomas (B). Propionate concentrations in fecal samples of individuals without colon tumors, adenomas, and carcinomas (C). The black lines indicate the median SCFA concentration. Acetate concentrations in fecal samples before and after treatment for adenoma (yellow) and carcinoma (red) (D). Butyrate concentrations in fecal samples before and after treatment for adenoma (yellow) and carcinoma (red) (E). Propionate concentrations in fecal samples before and after treatment for adenoma (yellow) and carcinoma (red) (F). The black dots and lines represent the median change in SCFA concentration.

Figure 2. No change in butyrate producing genes identified between normal, adenoma, and carcinoma individuals. Imputed gene relative abundance of important butyrate pathway genes using PICRUSt (A). Counts per million (corrected for size and number of contigs in an OPF) for the Butyrate Kinase gene (B). The other butyrate pathway genes from the PICRUSt analysis did not align to any of the OPFs in the metagenome analysis.

Figure 3. Patients with adenomas had the higest number of significant negative correlations between OTU relative abundance and SCFA concentration. Colors denote the family or lowest taxonomic ID that an OTU classified to. Fewer significant positive correlations were observed overall. Additionally, the differences in the number of significant positive correlations between patients with adenomas versus individuals without tumors (normal) and patients with carcinomas was not as pronounced as the number of significant negative correlations.

Figure 4. SCFA concentrations do not improve OTU-based Random Forest models. The area under the curve of 100 different 80/20 split OTU-based normal versus adenoma 10-fold CV models with and without SCFAs (A). The top 10 most important OTUs or SCFAs in the SCFA and OTU adenoma model (B). The top 10 most important OTUs in the OTU adenoma model (C). The area under the curve of 100 different 80/20 OTU-based normal versus carcinoma 10-fold CV models with and without SCFAs (D). The top 10 most important OTUs or SCFAs in the SCFA and OTU carcinoma model (E). The top 10 most important OTUs in the OTU carcinoma model (F). For

(A) and (D) the black line represents the median AUC. The dotted line highlights an AUC of 0.5.

Figure 5. OTU-based regression Random Forest models of SCFA concentrations. The train and test correlation between actual and predicted values from 100 different 80/20 split OTU-based models with 10-fold CV using regression Random Forest (A). The model accuracy of predicted SCFA concentrations differed between individuals without tumors, patients with adenomas, and patients with carcinomas. Generally, patients with carcinomas had predicted concentrations closest to their actual measured concentration (B). The top 10 OTUs based on mean decrease in accuracy (MDA) for each SCFA model, colored by their lowest taxonomic identification (C).

Figure S1. Patients with adenomas had the higest number of significant differences in OTU relative abundance between high/low SCFA groups. Colors denote the family or lowest taxonomic ID that an OTU classified to. Fewer significant OTUs were observed in individuals without tumors (normal) and patients with carcinomas versus patients with adenomas.

Figure S2. OTU-based classification Random Forest models of high/low SCFA groups based on overall SCFA median concentration. The train and test results of 100 different 80/20 split OTU-based models with 10-fold CV based on higher or lower than the median SCFA concentration using classification Random Forest (A). The model accuracy of predicted high/low SCFA groups differed between individuals without tumors, patients with adenomas, and patients with carcinomas. Patients with adenomas consistently had the best classification accuracy (B). The top 10 OTUs based on mean decrease in accuracy (MDA) for each SCFA model, colored by their lowest taxonomic identification (C).