Fecal short chain fatty acids are not predictive of colorectal cancer status and cannot be predicted based on bacterial community structure

Marc A. Sze[1], Begüm Topçuoğlu[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

Observation format

Abstract

The gut microbiome is thought to have a role in the development of colorectal cancer by protecting against and exacerbating inflammation. Short chain fatty acids (SCFAs), including butyrate, have been shown to have anti-inflammatory properties and are produced in large quantities by colonic bacteria. We assessed whether there was an association between fecal SCFA concentrations and the presence of colonic adenomas or carcinomas in a cohort of individuals that was previously subjected to 16S rRNA gene and metagenomic shotgun sequencing. We measured the fecal concentrations of acetate, propionate, isobutyrate, and butyrate within the cohort and found that there were no meaningful differences in their concentration and disease status. When we incorporated these concentrations into Random Forest classification models trained to differentiate between people with normal colons and those with adenomas or carcinomas, we found that they did not significantly improve the ability of 16S rRNA gene or metagenomic gene sequence-based models to classify individuals. Finally, we generated Random Forest regression models trained to predict the concentration of each SCFA based on 16S rRNA gene or metagenomic gene sequence data from the same samples. These models performed poorly and were only able to explain less than 10% of the observed variation in the SCFA concentrations. These results support the broader epidemiological data that questions the value of fiber consumption for reducing the risks of colorectal cancer. Although it is likely that bacterial metabolites may serve as biomarkers to detect adenomas or carcinomas, fecal SCFA concentrations have limited value.

Importance

Considering colorectal cancer is the third leading cancer-related cause of death within the United States, there is a great need to detect colorectal tumors early without invasive colonoscopy procedures and to prevent the formation of tumors. Short chain fatty acids (SCFAs) are often used as a surrogate for measuring gut health and for being anti-carcinogenic because of their anti-inflammatory properties. We evaluated the fecal SCFA concentration of a cohort of individuals with varying colonic tumor burden who were previously analyzed to identify microbiome-based biomarkers of tumors. We were unable to find an association between SCFA concentration and tumor burden or use SCFAs to improve our microbiome-based models of classifying people based on their tumor status. Furthermore, we were unable to find an association between the fecal community structure and SCFA concentrations. These data indicate that there is not a meaningful link between the gut microbiome, SCFAs, and tumor burden.

Colorectal cancer is the third leading cancer-related cause of death within the United States (1). Less than 10% of cases can be attributed to genetic risk factors (2). This leaves a significant role for environmental, behavioral, and other factors such as smoking and diet (3, 4). Colorectal cancer is thought to be initiated by a series of mutations that accumulate as mutated cells proliferate leading to adenomatous lesions, which are succeeded by carcinomas (2). Throughout this progression, there are ample opportunities for bacterial populations to create mutations, induce inflammation, and accelerate tumorigenesis. Numerous studies in murine models have supported this model (5–7). Additional cross sectional studies in humans have identified microbiome-based biomarkers of disease (8). These studies suggest that in some cases, it is the loss of bacterial populations that produce short-chain fatty acids (SCFAs) that results in increased inflammation and tumorigenesis.

SCFAs have have anti-inflammatory and anti-proliferative activities (9). Manipulation of SCFAs in mouse models of colorectal cancer by direct supplementation or feeding of fiber caused an overall reduction in tumor burden (10). These results suggest that supplementation with substrates that bacteria can ferment to produce SCFAs may confer beneficial effects against colorectal cancer. Regardless, there is a lack of evidence that increasing SCFA concentrations can protect against colorectal cancer in humans. Based on similar observations, many microbiome studies use the concentrations of SCFAs and the presence of 16S rRNA gene sequences from organisms and the genes involved in producing them as a biomarker of a healthy microbiota (11, 12). Case-control studies that have investigated SCFA concentrations in colorectal cancer found that patients with carcinomas had lower concentrations of SCFAs versus patients with adenomas or individuals without colon tumors (13). Although this would argue that increasing SCFA concentrations could be protective against tumorigenesis, in randomized controlled trials fiber supplementation has been inconsistently associated with protection against tumor formation and recurrence (14, 15). These findings temper enthusiasm for treatments that aim to use SCFAs as biomarkers or protection against tumorigenesis.

SCFA concentrations do not meaningfully vary with diagnosis or treatment. To quantify the associations between colorectal cancer, the microbiome, and SCFAs, we quantified the concentration of acetate, propionate, isobutyrate, and butyrate in feces of previously characterized individuals with normal colons (N=172) and those with colonic adenomas (N=198) or carcinomas

(N=120) (16). The only SCFA that had a significantly different concentration across the diagnoses was isobutyrate (P=0.0091; Figure 1A). Interestingly, the median concentration of isobutyrate was 3.30 mmol/kg in people with normal colons and it was 3.00 and 3.84 mmol/kg in people with colonic adenomas or carcinomas, respectively. The difference in isobutyrate concentration between people with adenomas and carcinomas was significantly different (P=0.0065); however, the differences between people with normal colons and those with adenomas or carcinomas was not significant (P=0.19 and P=0.11). Among the subjects with adenomas and carcinomas, a subset ($N_{adenoma}$=41, $N_{carcinoma}$=26) were treated and sampled a year later (17). The only SCFA that changed following treatment was isobutyrate, which decreased by 0.99 mmol/kg (P=0.002; Figure 1B). For both the pre-treatment cross-sectional data and the pre/post treatment data, we pooled the SCFA concentrations on a per molecule of carbon basis and again failed to see any significant differences (P>0.15). The low concentration of isobutyrate relative to the other SCFAs, inconsistent concentrations, and unexpected decrease in concentration with treatment makes it difficult to ascribe much biological relevance to this observation.

Combining SCFA and microbiome data does not improve the ability to diagnose individual as having adenomas or carcinomas. We previously found that binning 16S rRNA gene sequence data into operational taxonomic units based on 97% similarity or into genera enabled us to classify individuals as having adenomas or carcinomas using Random Forest machine learning models (8, 16). We repeated that analysis but added the concentration of the SCFAs as possible features to train the models (Figure S1). Models trained using SCFAs to classify individuals as having adenomas or carcinomas rather than normal colons had median areas under the receiver operator characteristic curve (AUROC) that were significantly greater than 0.5 ($P_{adenoma}$<0.001 and $P_{carcinoma}$<0.001); however, the AUROC values to detect the presence of adenomas or carcinomas were only 0.54 and 0.55, respectively (Figure 2A). When we trained the models with the SCFAs concentrations and operational taxonomic unit (OTU) or genus-level relative abundances the AUROC values were not significantly different from the models trained without the SCFA concentrations (P>0.21; Figure 2A). These data demonstrate that knowledge of the SCFA profile from a patient's fecal sample does not improve the ability to diagnose a colonic lesion.

Knowledge of microbial community structure does not predict SCFA concentrations. We next asked whether the fecal community structure was predictive of fecal SCFA concentrations, regardless of a person's diagnosis. We trained Random Forest regression models using 16S rRNA gene sequence data binned into OTUs and genera to predict the concentration of the SCFAs (Figure S2). The largest R^2 between the observed SCFA concentrations and the modeled concentrations was 0.14 observed when using genus data to predict butyrate concentrations (Figure 2B). We also used a smaller dataset of shotgun metagenomic sequencing data generated from a subset of our cohort (N_{normal}=27, $N_{adenoma}$=25, and N_{cancer}=26) (18). We binned genes extracted from the assembled metagenomes into operational protein families (OPFs) or KEGG categories and trained Random Forest regression models using metagenomic to predict the concentration of the SCFAs (Figure S2). Similar to the analysis using 16S rRNA gene sequence data, the metagenomic data was not predictive of SCFA concentration; the largest amount of variation that the models could explain was 0.055, which was observed when using KEGG data to predict propionate concentrations (Figure 2B). Because of the limited number of samples that we were able to generate metagenomic sequence data from, we used our 16S rRNA gene sequence data to impute metagenomes that were binned into metabolic pathways or KEGG categories using PICRUSt (Figure S2). SCFA concentrations could not be predicted based on the imputed metagenomic data. The largest amount of variation that the models could explain was 0.097 observed when using KEGG data to predict butyrate concentrations (Figure 2B). The inability to model SCFA concentrations from microbiome data indicates that the knowledge of the abundance of organisms and their genes was insufficient to predict SCFA concentrations.

Conclusion. Our data indicate that fecal SCFA concentrations are not associated with the presence of adenomas or carcinomas and that they provide weak predictive power to improve the ability to diagnose someone with one of these lesions. Furthermore, knowledge of the taxonomic and genetic structure of gut microbiota was not predictive SCFA concentrations. These results complement existing literature that suggest that fiber consumption and the production of SCFAs are unable to prevent the risk of developing colonic tumors. It is important to note that our analysis concerned fecal SCFA concentrations and microbiome characterization and that observations along the mucosa near the site of lesions may provide a stronger association. Regardless, given the growing literature in this area, it is unlikely that SCFAs are the primary mechanism that limits tumorigenesis. This may be a cautionary result to temper enthusiasm for SCFAs as a biomarker of gut health more generally. Going forward it is critical to develop additional hypotheses for how the microbiome and host interact to drive tumorigenesis to better understand the disease process and identify biomarkers that will allow early detection of tumors.

Materials and Methods

Study design and sampling. The overall study design and the resulting sequence data have been previously described (16, 17). In brief, fecal samples were obtained from 172 individuals with normal colons, 198 individuals with colonic adenomas, and 120 individuals with carcinomas. Of the individuals diagnosed as having adenomas or carcinomas, a subset ($N_{adenoma}$=41 and $N_{carcinoma}$=26) were sampled after treatment of the lesion (median=255 days between sampling, IQR=233 to 334 days). Tumor diagnosis was made by colonoscopic examination and histopathological review of the biopsies (16). The University of Michigan Institutional Review Board approved the studies that generated the samples and informed consent was obtained from all participants in accordance to the guidelines set out by the Helsinki Declaration.

Measuring specific SCFAs. The measurement of acetate, propionate, isobutyrate, and butyrate used a previously published protocol that used High-Performance Liquid Chromatography (HPLC) (19). Two changes were made to the protocol. First, instead of using fecal samples suspended in DNA Genotek OmniGut tubes, we suspended frozen fecal samples in 1 mL of PBS. Second, instead of using the average weight of fecal sample aliquots to normalize SCFA concentrations, we used the actual weight of the fecal samples. These methodological changes did not affect the range of concentrations of these SCFAs between the two studies.

16S rRNA gene sequence data analysis. Sequence data from Baxter et al. (16) and Sze et al. (17) were obtained from the Sequence Read Archive (studies SRP062005 and SRP096978) and reprocessed using using mothur v.1.42 (20). The original studies generated sequence data from V4 region of the 16S rRNA gene using paired 250 nt reads on an Illumina MiSeq sequencer. The resulting sequence data were assembled into contigs, screened to remove low quality contigs and chimeras. The curated sequences were then clustered into OTUs at a 97% similarity threshold and assigned to the closest possible genus with an 80% confidence threshold trained on the reference collection from the Ribosomal Database Project (v.16). We used PICRUSt (v.2.1.0-b) with the recommended standard operating protocol to generate imputed metagenomes based on the expected metabolic pathways, KEGG categories, and enzyme commission numbers (21).

Metagenomic DNA sequence analysis. A subset of the samples from the samples described by Baxter et al. (16) were used to generate metagenomic sequence data (N_{normal}=27, $N_{adenoma}$=25, and N_{cancer}=26). These data were generated by Hannigan et al. (18) and deposited into the Sequence Read Archive (study SRP108915). Fecal DNA was subjected to shotgun sequencing on an Illumina HiSeq using 125 bp paired end reads. The archived sequences were already quality filtered and aligned to the human genome to remove contaminating sequence data. We downloaded the sequences and assembled them into contigs using MEGAHIT (22), which were used to identify open reading frames (ORFs) using Prodigal (23). We determined the abundance of each ORF by mapping the raw reads back to the ORFs using Diamond (24). We clustered the ORFs into operational protein families (OPFs) in which the clustered ORFs were more than 40% identical to each other using mmseq2 (25). We also used mmseq2 to map the ORFs to the KEGG database and clustered the ORFs according to which category the ORFs mapped.

Random Forest models. The classification models were built to predict lesion type from microbiome information with or without SCFA concentrations. The regression models were built to predict the SCFA concentrations of acetate, butyrate, and propionate from microbiome information. For classification and regression models, we pre-processed the features by scaling them to vary between zero and one. Features with no variance in the training set were removed from both the training and testing sets. We randomly split the data into training and test sets so that the training set consisted of 80% of the full dataset while the test set was composed of the remaining data. The training set was used for hyperparameter selection and training the model, and the test set was used for evaluating prediction performance. For each model, the best performing hyperparameter, mtry, was selected in an internal five-fold cross-validation of the training set with 100 randomizations. Six values of mtry were tested and the value that provided the largest AUROC or R^2 was selected. We trained the random forest model using the selected mtry value and predicted the held-out test set. The data-split, hyperparameter selection, training and testing steps were repeated 100 times to get a reliable and robust reading of model prediction performance. We used AUROC and R^2 as the prediction performance metric for classification and regression models, respectively. We used randomForest package implemented to the caret package (version 4.6-14) in R statistical software (version 6.0-81) for our models.

Statistical analysis workflow. Data summaries, statistical analysis, and data visualizations were performed using R (v.3.5.1) with the tidyverse package (v.1.2.1). To assess differences in SCFA concentrations between individuals normal colons and those with adenomas or carcinomas, we used the Kruskal-Wallis rank sum test. If a test had a P-value below 0.05, we then applied a pairwise Wilcoxon rank sum test with a Benjamini-Hochberg correction for multiple comparisons. To assess differences in SCFA concentrations between individuals samples before and after treatment we used paired Wilcoxon rank sum tests to test for significance. To compare the median AUCROC for the held out data for the model generated using only the SCFAs, we compared the distribution of the data to the expected median of 0.5 using the Wilcoxon rank sum test to test whether the model performed better than would be achieved by randomly assigning the data to each diagnosis. When we compared the Random Forest models generated without and with SCFA data included, we used Wilcoxon rank sum tests to determine whether the models with the SCFA data included did better.

Code availability. The code for all sequence curation and analysis steps including an Rmarkdown version of this manuscript is available at https://github.com/SchlossLab/Sze_SCFACRC_XXXX_2019/.

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 thank the University of Michigan Center for Microbial Systems for enabling our short-chain fatty acid analysis. Support for MAS came from the Canadian Institute of Health Research and the National Institutes of Health (UL1TR002240). Support for PDS came from the National Institutes of Health (P30DK034933 and R01CA215574).

References

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

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

3. **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.

4. **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.

5. **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.

6. **Shields CED**, **Meerbeke SWV**, **Housseau F**, **Wang H**, **Huso DL**, **Casero RA**, **O'Hagan HM**, **Sears CL**. 2016. Reduction of murine colon tumorigenesis driven by Enterotoxigenic Bacteroides fragilis using cefoxitin treatment. Journal of Infectious Diseases **214**:122–129. doi:10.1093/infdis/jiw069.

7. **Tomkovich S**, **Yang Y**, **Winglee K**, **Gauthier J**, **Mühlbauer M**, **Sun X**, **Mohamadzadeh M**, **Liu X**, **Martin P**, **Wang GP**, **Oswald E**, **Fodor AA**, **Jobin C**. 2017. Locoregional effects of microbiota in a preclinical model of colon carcinogenesis. Cancer Research **77**:2620–2632. doi:10.1158/0008-5472.can-16-3472.

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

9. **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.

10. **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.

11. **Sanna S**, **Zuydam NR van**, **Mahajan A**, **Kurilshikov A**, **Vila AV**, **Võsa U**, **Mujagic Z**, **Masclee AAM**, **Jonkers DMAE**, **Oosting M**, **Joosten LAB**, **Netea MG**, **Franke L**, **Zhernakova A**, **Fu J**, **Wijmenga C**, **McCarthy MI**. 2019. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. Nature Genetics. doi:10.1038/s41588-019-0350-x.

12. **Meisel M**, **Mayassi T**, **Fehlner-Peach H**, **Koval JC**, **O'Brien SL**, **Hinterleitner R**, **Lesko K**, **Kim S**, **Bouziat R**, **Chen L**, **Weber CR**, **Mazmanian SK**, **Jabri B**, **Antonopoulos DA**. 2016. Interleukin-15 promotes intestinal dysbiosis with butyrate deficiency associated with increased susceptibility to colitis. The ISME Journal **11**:15–30. doi:10.1038/ismej.2016.114.

13. **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.

14. **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.

15. **Gianfredi V**, **Salvatori T**, **Villarini M**, **Moretti M**, **Nucci D**, **Realdon S**. 2018. Is dietary fibre truly protective against colon cancer? A systematic review and meta-analysis. International Journal of Food Sciences and Nutrition **69**:904–915. doi:10.1080/09637486.2018.1446917.

16. **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.

17. **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.

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

19. **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.

20. **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.

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

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

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

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

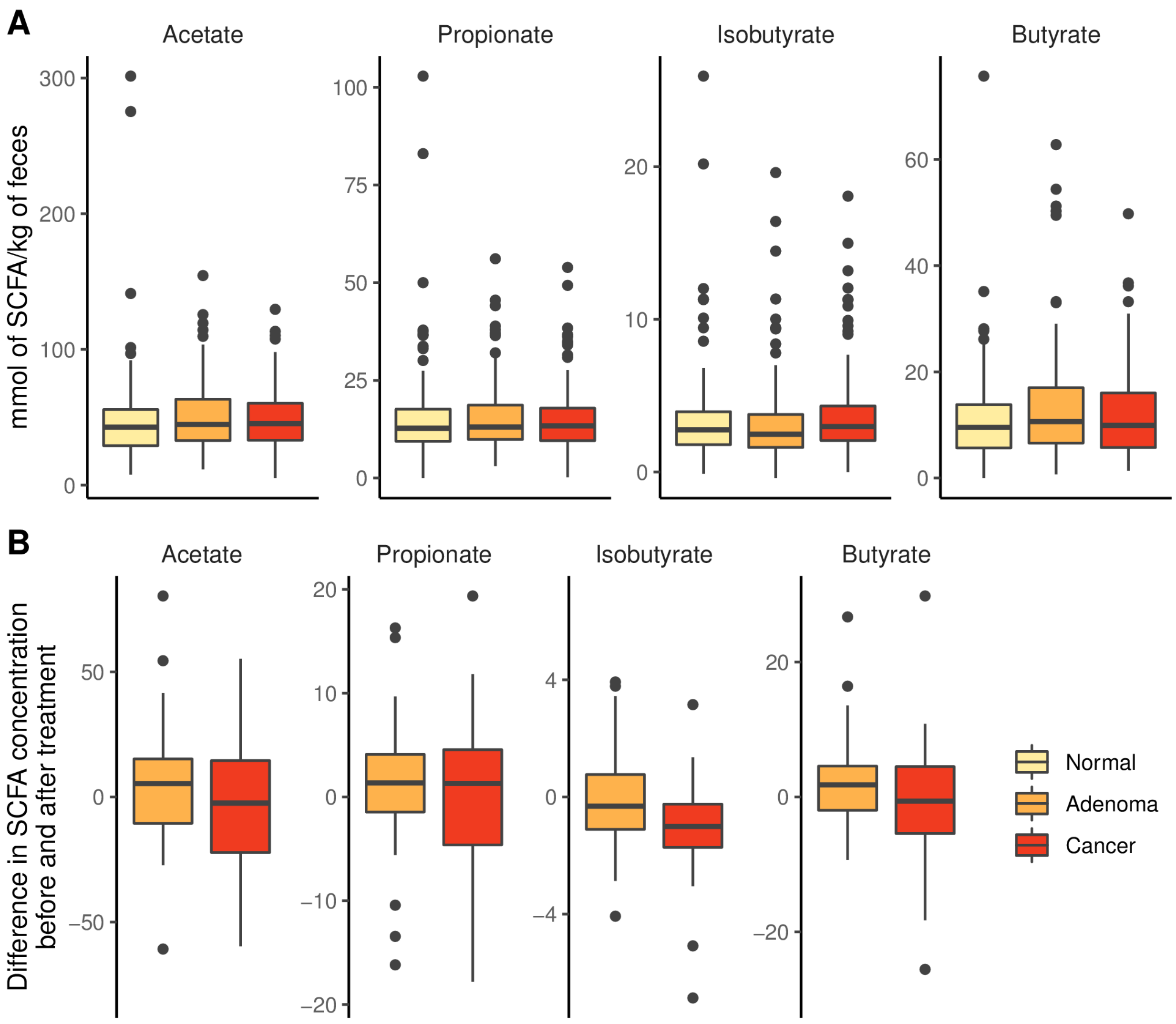

Figure 1. SCFA concentrations did not vary meaningfully with diagnosis of colonic lesions or with treatment for adenomas or carcinomas. (A) We measured the concentration of fecal SCFAs from individuals with normal colons (N=172) or those with adenoma (N=198) or carcinomas (N=120). (B) A subset of individuals diagnosed with adenomas (N=41) or carcinomas (N=26) who underwent treatment were resampled a year after the initial sampling; one extreme propionate value (124.4 mmol/kg) was included in the adenoma analysis but censored from the visualization for clarity.

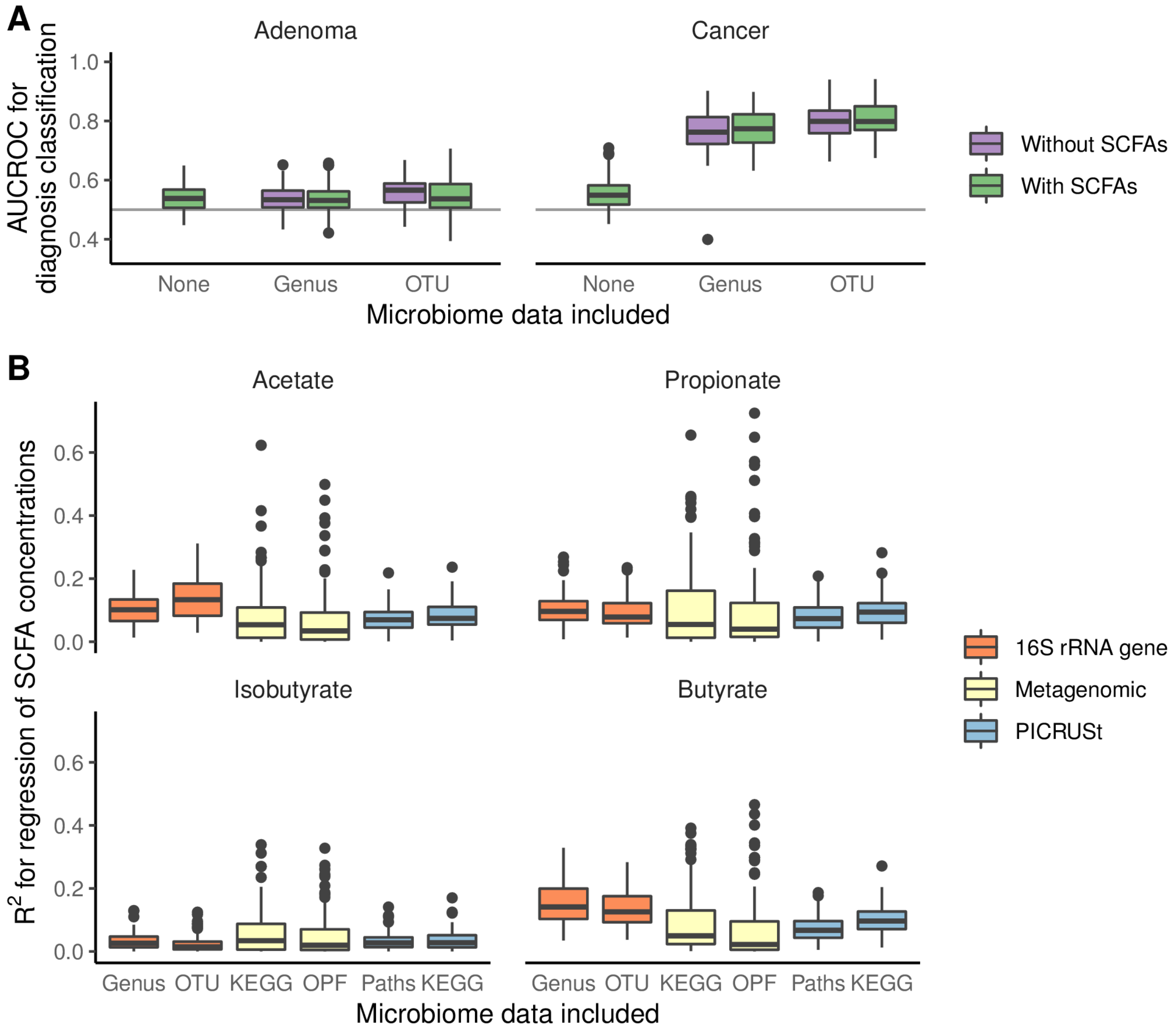

Figure 2. SCFA concentrations do not improve models for diagnosing the presence of adenomas, carcinomas, or all lesions and cannot be reliably predicted from 16S rRNA gene or metagenomic sequence data. (A) The median AUROC for diagnosing individuals as having adenomas or carcinomas using SCFAs was slightly better than than chance (depicted by horizontal line at 0.50), but did improve performance of the models generated using 16S rRNA gene sequence data. (B) Regression models that were trained using 16S rRNA gene sequence, metagenomic, and PICRUSt data to predict the concentrations of SCFAs performed poorly (all median R^2 values < 0.14). Regression models generated using 16S rRNA gene sequence and PICRUSt data included data from 490 samples and those generated using metagenomic data included data from 78 samples.

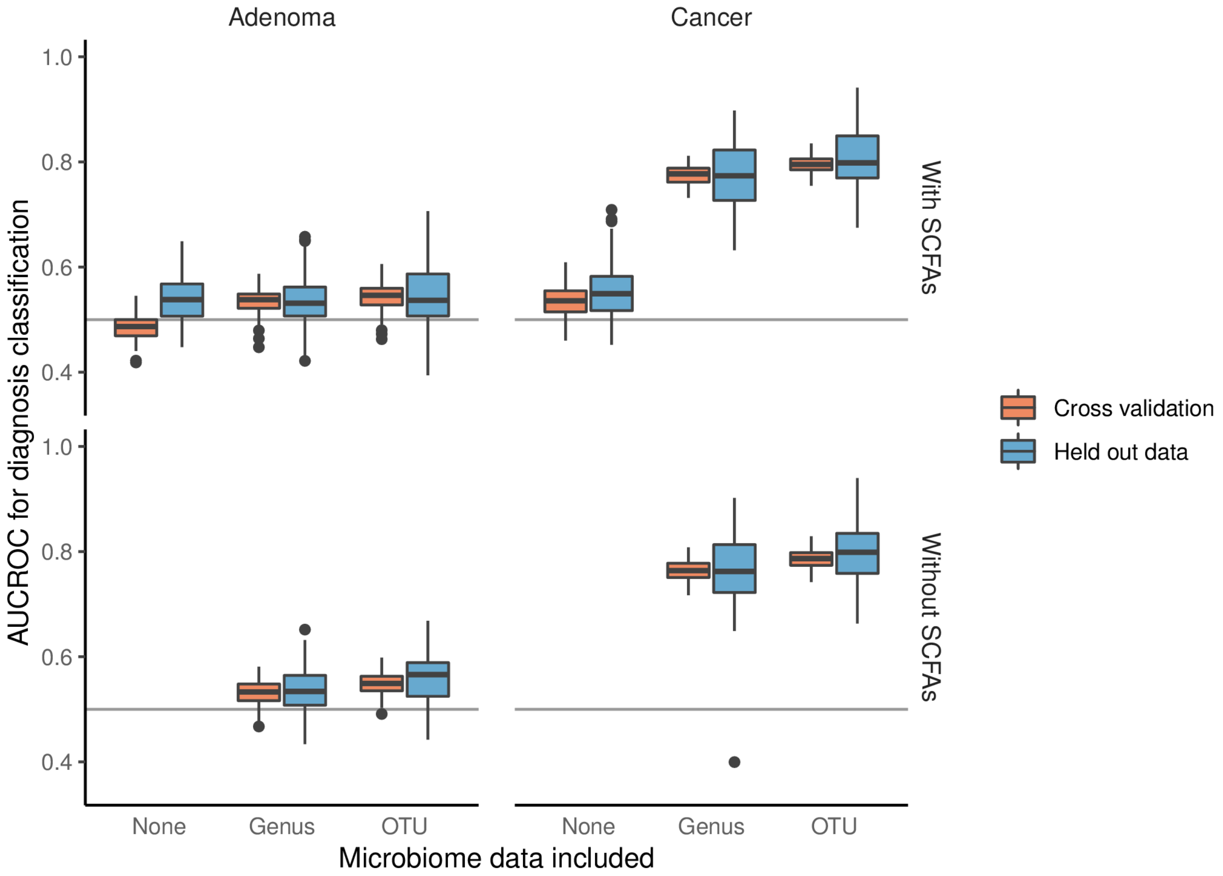

Figure S1. Comparison of training and testing results for classification models shows that the models are robust and are not overfit. Random Forest classification models were generated to differentiate between subjects with normal colons and those with adenomas or carcinomas using 16S rRNA gene sequence data that were clustered into genera or OTUs with and without including the four SCFAs as additional features. Random Forest classification models were generated by partitioning the samples into a training set with 80% of the data and a testing set with the remaining samples. The training data was used to fit the mtry hyperparameter using 100 five-fold cross validations for six values of mtry. The mtry with the largest AUROC was selected to train the model, which was then tested on the held out 20% of the samples.

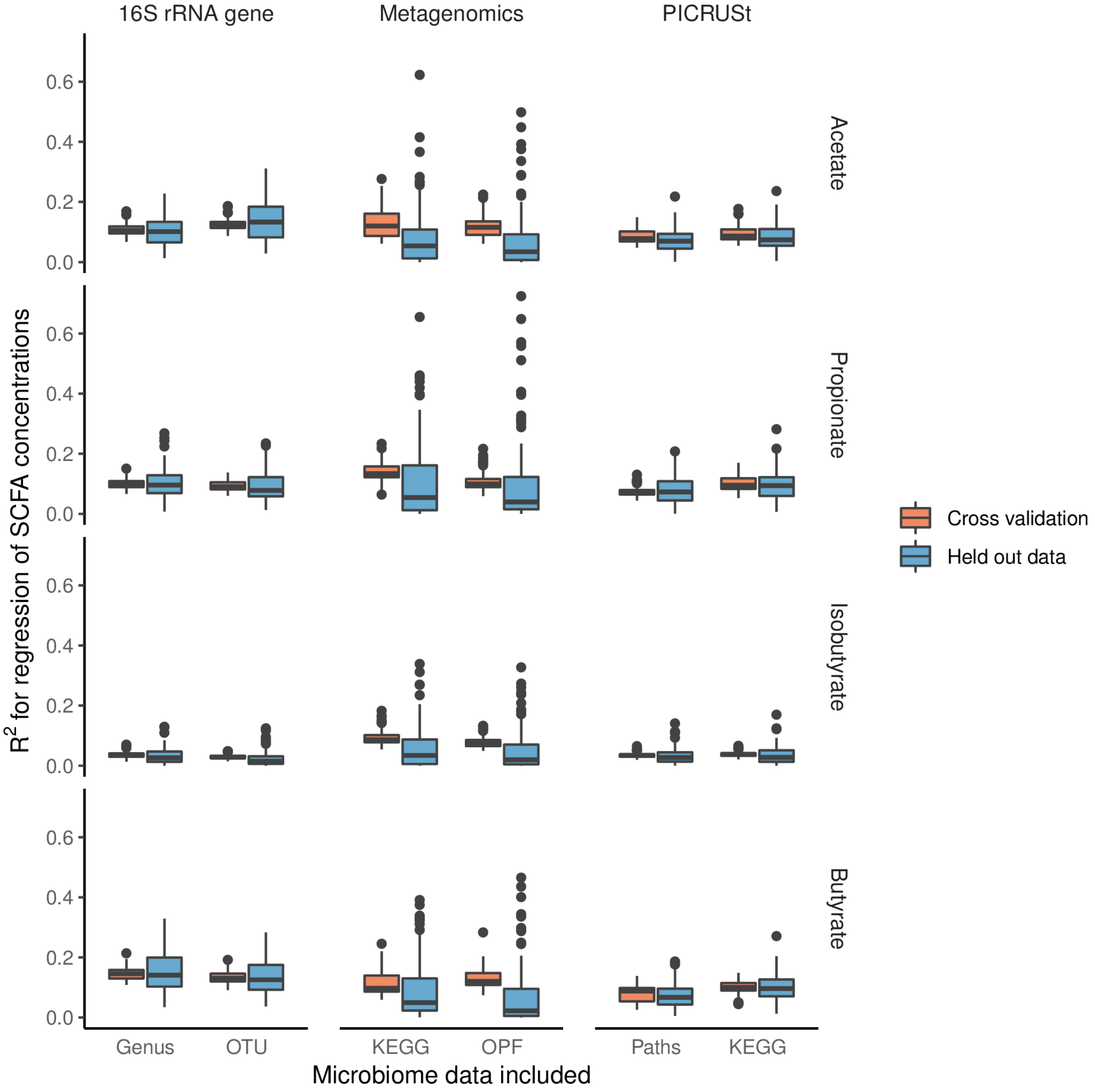

Figure S2. Comparison of training and testing results for regression models shows that the models are robust and are not overfit. Random Forest regression models were generated to predict the concentration of each SCFA using each subjects' microbiome data generated using 16S rRNA gene sequence and metagenomic sequence data. These regression models were generated by partitioning the samples into a training set with 80% of the data and a testing set with the remaining samples. The training data was used to fit the mtry hyperparameter using 100 five-fold cross validations for six values of mtry. The mtry with the largest R^2 was selected to train

the model, which was then tested on the held out 20% of the samples.