Revisiting Short-Chain Fatty Acids and the Microbiota in Colorectal Cancer

Running title: SCFAs and Colorectal Cancer

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

Introduction

Results

Decreased SCFA concentrations are not associated with tumors. We used frozen fecal samples from 490 individuals and HPLC to measure acetate, butyrate, and propionate concentrations at a cross-sectional point in time. There was no difference between individuals with normal colons and patients with either an adenoma or carcinoma 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 or carcinoma in which we had pre- and post-treatment fecal samples. Although there was a general trend for increasing acetate, butyrate, and propionate concentrations after treatment for tumors, there was no significant difference pre- and post-treatment for patients with adenoma (P-value > 0.20) or carcinoma (P-value > 0.80) [Figure 1D - 1F]. Even though there was no change in SCFA concentrations between individuals with normal colons and those with tumors, this information could still be important to help classify disease.

Random Forest models with SCFA concentrations do not classify tumor better. Using the Random Forest algorithm we built models to classify normal versus adenoma and normal versus carcinoma with OTUs or OTUs and SCFA concentrations. For both adenoma and carcinoma models, there was no difference between the median AUC of models with or without SCFA concentrations (P-value > 0.05) [Figure 2]. Although there was no added information gained for classification of disease by including SCFAs to Random Forest models, it is possible that the genes for enzymes involved in SCFA pathways may change due to disease status; where a smaller number of microbes are responsible for the observed SCFA concentrations.

Discussion

Conclusions

Materials and Methods

Study design and sampling. The overall protocol has been described in detail previously (1, 2). 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 for their tumor (n=67). 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 (1). Diagnosis of tumor was made by colonoscopic examination and histopathological review of biopsies obtained (1, 2). 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 Decleration.

Measuring specific SCFAs. Our protocol for the measurement of acetate, butyrate, and propionate followed a previously published protocol (3). 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 acutal weight of fecal samples instead of the average weight for SCFA concentration normalizations. These changes did not affect the overall median concentrations of these SCFAs between the two studies (see Table 1 (3) and Figure 1 in this report).

16s rRNA gene sequencing. The workflow and processing have been described previously (1, 4, 5). 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 used the OptClust algorithm (6).

Generating imputed metagenomes. The use of PICRUSt version 1.1.2 with the recommended standard operating protocol (7) was used. Briefly, the mothur shared file and metadata was converted into a biom formated 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 OPFs from metagenomes. A subset of the cross-sectional group (n=490) containing a total of 85 individuals (normal n=29 normal, 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 (8). Briefly, the sequences were quality filtered and sequences aligning to the human genome were removed prior to contig assembly with MEGAHIT (9). Open Reading Frames (ORFs) were identified using Prodigal (10), counts generated using Diamond (11), subsequent clustering into Operational Protein Families (OPFs) used mmseq2 (12), and OPF alignment used the KEGG database (13).

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 (13, 14) 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 both classification and regression models (15). This was repeated on 100 differen 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. Classification models were built to group normal versus adenoma and normal versus carcinoma or high versus low SCFA concentrations. Regression models were built to assess how well OTUs could be used to get the SCFA concentration within a sample.

Statistical analysis workflow. All analysis was performed using the statistical language R (16). Generally, differences between the different disease groups used a Kruskal-Walis rank sum test with a Dunn's post-hoc test. Where appropriate Benjamini-Hochberg was used to correct for multiple comparisons (17). 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, we assessed whether OTUs alone or OTUs and SCFAs were better able to classify individuals with and without tumor using Random Forest models. Next, the imputed gene counts of important mediators of SCFA creation were tested. Finally, the counts generated for OPFs that matched important genes involved with SCFA creation were analyzed. Finally, models to classify high or low SCFA concentration based on 16S rRNA gene sequencing data were created using Random Forest. Regression models also were created to classify the exact SCFA concentration based on 16S rRNA gene sequencing data also were built using the Random Forest algorithm. This was done to assess the ability of 16S rRNA gene sequencing to classify the SCFA concentrations.

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-chian fatty acid analysis on the High-Performace 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. **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.

2. **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.

3. **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.

4. **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.

5. **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.

6. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere **2**:e00073–17. doi:10.1128/mspheredirect.00073-17.

7. **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.

8. **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.

9. **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.

10. **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.

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

12. **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.

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

14. **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.

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

16. **R Core Team**. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

17. **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. Using HPLC no change in SCFA measurements was observed between normal, adenoma, and carcinoma individuals. Acetate concentrations in fecal samples of individuals with normal colons, adenomas, and carcinomas (A). Butyrate concentrations in fecal samples of individuals with normal colons, adenomas, and carcinomas (B). Propionate concentrations in fecal samples of individuals with normal colons, adenomas, and carcinomas (C). The black links 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. SCFAs do not improve OTU-based Random Forest models. Difference between 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). Difference between the area under the curve of 100 different 80/20 OTU-based normal versus carcinoma 10-fold CV models with and without SCFAs (B). The black linke represents the median AUC. The dotted line highlights an AUC of 0.5.

Figure 3. 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 genes from the PICRUSt analysis did not align to any of the OPFs in the metagenome analysis.

Figure S1. OTU-based Random Forest models of SCFA concentrations. Classification Random Forest train and tests of 100 different 80/20 OTU-based models with 10-fold CV based on higher or lower than the medain SCFA concentration (A). The top 10 OTUs based on mean decrease in accuracy (MDA) for each model, colored by their lowest taxonomic identification (B). Regression Random Forest train and tests of 100 different 80/20 OTU-based models with 10-fold CV based on correlation to actual SCFA concentration (C). The top 10 OTUs based on mean decrease in accuracy (MDA) for each model, colored by their lowest taxonomic identification (D).