

The Fecal Microbiome Before and After Treatment for Colorectal Adenoma or Carcinoma

Running Title: Human Microbiome before and after Colorectal Cancer

Marc A Sze¹, Nielson T Baxter², Mack T Ruffin IV³, Mary AM Rogers², and 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

² Department of Internal Medicine, University of Michigan, Ann Arbor, MI

³ Department of Family Medicine and Community Medicine, Penn State Hershey Medical Center, Hershey, PA

Abstract

Background: Colorectal cancer (CRC) is a worldwide health problem and research suggests a correlation between the fecal bacterial microbiome and CRC. This study tested the hypothesis that treatment for adenoma or carcinoma result in changes to the bacterial community. Specifically, we tried to identify components within the community that were different before and after removal of lesion (adenoma or carcinoma).

Results: There was a larger change in the bacterial community in response to treatment for carcinoma versus adenoma cases (P -value < 0.05). Yet no difference was found in the relative abundance of any OTU before and after treatment for adenoma or carcinoma groups (P -value > 0.05). A lesion model had an AUC range of 0.692 - 0.761 and follow up samples had no difference in the positive probability of lesion versus initial samples (P -value > 0.05); suggesting that the lesion associated community persists after treatment. A treatment model had an AUC range of 0.657 - 0.796 and had a decrease positive probability for the follow up samples to be an initial sample (P -value < 0.05); suggesting that there are members within the community that repond to treatment. The lesion model used a total of 54 variables while the initial sample model used a total of 70 variables. A total of 32 OTUs were common to both models with many classifying to commensal bacteria (e.g. *Lachnospiraceae*, *Bacteroides*, *Roseburia*, *Blautia*, and *Ruminococcus*).

Conclusions: Our data partially supports the hypothesis that the bacterial community changes after treatment. Individuals with carcinoma have more drastic differences to the overall community then those with adenoma. Commensal bacteria were crucial for accurate model classification, suggesting that these bacteria may be important to initial polyp formation and transition to carcinoma.

24 **Keywords**

25 bacterial microbiome; colorectal cancer; polyps; FIT; post-surgery; risk factors

Background

Colorectal cancer (CRC) is currently the third most common cause of cancer deaths [1,2]. The rate of disease mortality has seen a significant decrease, thanks mainly to improvements in screening [1]. However, despite this improvement there are still approximately 50,000 deaths from the disease per year [2].

Recent studies in humans have shown that both the bacterial community and specific members within it correlate with CRC pathogenesis [3,4]. Further, Dejea, et al. observed that bacterial communities are altered between normal and tumor tissue [5]. Mouse models of CRC have further demonstrated the importance of the microbiome, both on a community [3,6] and species level [4], for tumorigenesis. Collectively, these studies provide a tantalizing link between our gut bacteria and CRC and suggest that biomarkers using our microbes could be developed. Indeed, building models using 16S rRNA gene sequencing along with clinical tests such as Fecal Immunoglobulin Test (FIT) result in good predictions of CRC [7,8]. Although these studies show how our gut bacteria can impact CRC progression via a changed community or invasion by more inflammatory bacteria [9]. They provide very little information as to whether these changed communities rebound after successful treatment of lesion (adenoma or carcinoma).

In this study, we tested the hypothesis that there are detectable changes to the bacterial community between pre- (initial) and post- (follow up) treatment of lesion. We analyzed changes in alpha and beta diversity as well as the relative abundance of specific Operational Taxonomic Units (OTUs). We then investigated several important genera from oral microbes that have been suggested to be important in CRC pathogenesis [9]. We next utilized Random Forest to build two models: The first was built to classify lesion versus non-lesion (normal) while the second was built to classify initial versus follow up treatment samples. Subsequent observations on how these models and OTUs within them

51 performed before and after treatment helped inform us as to whether they were changing
52 after treatment and whether it was towards a normal community. We also investigated
53 the two models for similar important OTUs to identify which were important for classifying
54 both lesion versus normal and initial versus follow up treatment samples. Finally, since
55 treatment also included chemotherapy and radiation therapy along with lesion removal, we
56 tested if these additionally treatments significantly lowered any of the metrics examined
57 versus those who received only lesion removal. This study helps to provide evidence as to
58 whether treatment can influence the community and if the CRC microbiome, identified in
59 previous studies, persists.

Results

The Bacterial Community: Within our 67-person cohort we tested whether those with adenoma (n = 41) or carcinoma (n = 26) had any broad differences between their initial and follow up samples. We found that those with carcinoma had a more dissimilar bacterial community between their initial and follow up sample than those with adenoma (P-value < 0.001) [Figure 1A]. The bacterial community structure before and after surgery was visualized using NMDS for both adenoma [Figure 1B] (PERMANOVA > 0.05) and carcinoma [Figure 1C] (PERMANOVA < 0.05). Interestingly, when initial and follow up samples were compared, regardless of whether the lesions were adenoma or carcinoma, there was no significant overall difference in beta diversity (PERMANOVA > 0.05). There was no difference between initial and follow up samples when investigating alpha diversity metrics for lesion, adenoma only, or carcinoma only for any metric tested [Table S1]. Additionally, there was also no difference in the relative abundance of any OTU between initial and follow up samples for lesion, adenoma only, or carcinoma only [Figure S1].

Carcinoma Associated Genera: For carcinoma [Figure S2A] (P-value > 0.05) but not adenoma [Figure S2B] (P-value > 0.05) there was a significant decrease in *Porphyromonas*, *Parvimonas*, and *Peptostreptococcus* between initial and follow up samples [Figure S2]. *Fusobacterium* for both adenoma and carcinoma had no significant differences between initial and follow up samples [Figure S2] (P-value > 0.05). Although there were significant differences, only a small percentage of those with adenoma or carcinoma were positive for any of these specific genera [Figure S2].

The Lesion Model: We tested whether follow up samples were more normal than initial samples by building a model based solely on OTUs alone that classified lesion versus normal. This model had an AUC range of 0.692 - 0.761 after 100 iterations of 20 repeated 10-fold cross validations. The ROC curve for the final lesion model used was within the

observed range of the 100-different test set AUC iterations [Figure 2A]. There was a total of 54 OTUs that were used in this model [Figure 2B]. *Lachnospiraceae* (Otu000015) resulted in the largest decrease in MDA [Figure 2B]. Heavily represented genera within the model included OTUs from *Lachnospiraceae*, *Bacteroides*, *Roseburia*, *Blautia*, and *Ruminococcus*.

If there was movement towards a more normal community after treatment of lesion, we would expect to find a decrease in the positive probability of the follow up sample to be a lesion. We observed no such decrease in positive probability for either those with adenoma or carcinoma [Figure 2C & 2D] (P-value > 0.05). We also observed a significant difference between the predicted and actual calls (P-value < 0.05). However, this model correctly classified the one individual who still had carcinoma on follow up and this individual's positive probability of lesion increased between their initial and follow up sample [Figure 2C].

The Treatment Model: To test if there were differences that could be used to classify initial and follow up treatment samples we built a model to classify initial samples. This treatment model had an AUC range of 0.657 to 0.796. after 100 iterations of 20 repeated 10-fold cross validations. The test set AUC range for this model performed better than the training set AUCs. There was a marked decrease in the ROC curve for the final model used when compared to the 100 test set AUC iterations [Figure 3A]. There was a total of 70 OTUs that were used for this model [Figure 3B]. The variable that resulted in the largest MDA was *Ruminococcaceae* (Otu000278) [Figure 3B]. Similar to the lesion model, heavily represented genera included OTUs from *Lachnospiraceae*, *Bacteroides*, *Roseburia*, *Blautia*, and *Ruminococcus*.

If there were changes between initial and follow up samples we would expect the positive probability of being an initial sample to be decreased in the follow up. This is what we observed for the treatment model (P-value < 0.001). When we separated lesion into

adenoma and carcinoma there was a decrease in positive probability for both the carcinoma [Figure 3C] (P-value < 0.001) and adenoma group [Figure 3D] < 0.001). For this model, there was no difference between the predicted and actual classifications (P-value > 0.05).

Common OTUs to both Models: We next wanted to know what predictors within the lesion model were also in the initial sample model. The main purpose was to identify which OTUs could be important in lesion formation and were impacted by treatment. When we compared the two different models with each other there were a total of 32 common OTUs. Some of the most common taxonomic identifications belonged to *Lachnospiraceae*, *Bacteroides*, *Roseburia*, *Blautia*, *Anaerostipes*, *Dorea*, and *Ruminococcus*. These along with the vast majority of the OTUs that were common between models had classifications to bacteria typically thought of as commensal [Table S2].

Chemotherapy and Radiation Differences: After observing these changes due to treatment we assessed whether chemotherapy or radiation impacted the observed results. Only the treatment model had a significant change in positive probability for those treated with chemotherapy or radiation versus those who received neither [Figure 4A & Table S3] (P-value < 0.05). Additionally, there was no difference between the change in positive probability for chemotherapy and radiation therapy (Uncorrected P-value > 0.05). Using the 32 common OTUs we tested how chemotherapy or radiation affects important members of the community. For both chemotherapy and radiation, the directionality of change of these OTUs were similar [Figure 4B & 4C]. After multiple comparison correction, only *Blautia* (Otu000006) remained significantly different between those who did and did not receive chemotherapy. This data suggests that follow up samples from those treated with either chemotherapy or radiation may have had a larger change from initial sample due to effects on the bacterial community.

Discussion

This study builds upon previous work from numerous labs that have considered both how the bacterial community between those with and without CRC differ and how it might be used as an early screening tool [7,8,10–12]. Here we describe how the bacterial community changes after treatment and which OTUs may be most important to lesion classification. Interestingly, many of the most important OTUs for both treatment and lesion classification had taxonomic identification for resident gut microbes. This suggest that members within the commensal community may be the first that change during CRC pathogenesis. These changes, in turn, could be the first step in allowing more inflammatory bacteria to gain a foothold within the colon [9].

Unlike previous studies on the microbiome and CRC, ours focuses on identifying commonalities within both adenoma and carcinoma groups before and after treatment. Although there were differences for genera associated with specific bacterium linked with CRC [Figure S2]. These changes were not consistent across lesion. Instead, most of our results provide support that the first members of the community to change and potentially stay changed even after treatment are those that are commensal bacteria [Figure 2-3 & Table S3]. Additionally, treatment with chemotherapy or radiation changed these bacteria and provided a larger decrease in positive probability then removal of lesion.

Curiously, we observed that the typical CRC associated bacteria were not predictive within our models. There are several reasons why this may have occurred. First, is that they were not present in enough individuals to be able to classify those with and without disease with a high degree of accuracy. Second, is that our Random Forest models were able to gather the same information from other OTUs. Third, is that changes in commensal bacteria are the common differences across adenoma and carcinoma. Finally, it is also possible that all these explanations could have played a role. Our observations though,

would suggest that an individual's resident bacteria have a large role to play in polyp formation and could change in a way that allows predictive models to lower the positive probability of lesion or initial sample before treatment [Figure 2 & 3]. It should be noted that our study does not argue against the importance of these CRC associated bacteria in the pathogenesis of disease but rather that they are not the main bacteria changing after treatment. It is possible that these CRC associated bacteria are important in the transition from adenoma to carcinoma and would be one explanation as to why in our data we not only see high initial relative abundances in carcinoma and not adenoma individuals but also large decreases in relative abundance in some of those with carcinoma but not in those with adenoma after treatment [Figure S2].

Many of the common OTUs that we identified taxonomically classified to potential butyrate producers [Table S2]. Another batch of OTUs classified to bacteria that can either degrade polyphenols or are inhibited by them. Both butyrate and polyphenols are thought to be protective against cancer in part by reducing inflammation [13]. These protective compounds are derived from the breakdown of fiber, fruits, and vegetables by resident gut microbes. One example of this potential diet-microbiome-inflammation-polyp axis is that *Bacteroides*, which was highly prevalent in our models, are known to be increased in those with high non-meat based protein consumption [14]. High protein consumption in general has been linked with an increased CRC risk [15]. Conversely, *Bacteroides* are inhibited by polyphenols which are derived from fruits and vegetables [16]. Our data fits with the hypothesis that the microbial metabolites from breakdown products within our own diet could not only help to shape the existing community but also have an effect on CRC risk and disease progression.

A limitation, in our study, was that there was a significant difference in the time elapsed in the collection of the follow up sample between adenoma and carcinoma (uncorrected P-value < 0.05), with time passed being less for adenoma (253 +/- 41.3 days) than

carcinoma (351 +/- 102 days). These results would indicate that some of the differences observed between the carcinoma and adenoma groups could be due to differences in collection time. Specifically, it could confound the observation that carcinomas changed more than adenomas [Figure 1A & 1C]. This confounding though would not affect the observations where these individuals were grouped together [Figures 2-4].

Another limitation was that we do not know whether individuals who were still classified as positive by the lesion model eventually had a subsequent CRC diagnosis. This information would help to strengthen the case for this model keeping several individuals above the cutoff threshold even though at follow up they were diagnosed as no longer having a lesion. This study also drew heavily from those with Caucasian ancestry making it possible that the observations may not be representative of those with either Asian or African ancestry. Although our training and test set are relatively large we still run the risk of over-fitting or having a model that may not be representative of other populations. We've done our best to safeguard against this by not only running 10-fold cross validation but also having over 100 different 80/20 splits to try and mimic the type of variation that might be expected to occur.

Interestingly, within the treatment model the test data performed better than the training data. This may have occurred because the training AUC determined from 20 repeated 10-fold cross validation removed samples at random and did not consider that they were matched samples. Another potential reason is that the model itself may have been over-fit since the total number of samples was not that large. However, the lesion model did not suffer from these discrepancies. Further independent studies need to be carried out to verify our findings on lesion and treatment changes due to these limitations.

Despite these shortcomings our findings add to the existing scientific knowledge on CRC and the microbiome: That there is a measurable difference in the bacterial community after adenoma and carcinoma treatment. Further, the ability for machine learning algorithms

212 to take OTU data and successfully lower positive probability of lesion or initial sample
213 after treatment provides evidence that there are specific signatures, mostly attributable
214 to commensal organisms, associated with both treatment and lesion. Our data provides
215 evidence that commensal bacteria may be important in the development of polyps and
216 potentially the transition from adenoma to carcinoma.

Methods

Study Design and Patient Sampling: Sampling and design have been previously reported in Baxter, et al [7]. Briefly, study exclusion involved those who had already undergone surgery, radiation, or chemotherapy, had colorectal cancer before a baseline fecal sample could be obtained, had IBD, a known hereditary non-polyposis colorectal cancer, or familial adenomatous polyposis. Samples used to build the models for prediction were collected either prior to a colonoscopy or between 1 - 2 weeks after. The bacterial community has been shown to normalize back to a pre-colonoscopy community within this time period [17]. Our study cohort consisted of 67 individuals with an initial sample as described and a follow up sample obtained between 188 - 546 days after treatment of lesion. This study was approved by the University of Michigan Institutional Review Board. All study participants provided informed consent and the study itself conformed to the guidelines set out by the Helsinki Declaration.

16S rRNA Gene Sequencing: Sequencing was completed as described by Kozich, et al. [18]. DNA extraction used the 96-well Soil DNA isolation kit (MO BIO Laboratories) and an epMotion 5075 automated pipetting system (Eppendorf). The V4 variable region was amplified and the resulting product was split between three sequencing runs with normal, adenoma, and carcinoma evenly represented on each run. Each group was randomly assigned to avoid biases based on sample collection location. The initial and follow up samples were sequenced on the same run.

Sequence Processing: The mothur software package (v1.37.5) was used to process the 16S rRNA gene sequences and has been previously described [18]. The general workflow using mothur was: Paired-end reads were first merged into contigs, quality filtered, aligned to the SILVA database, screened for chimeras, classified with a naive Bayesian classifier using the Ribosomal Database Project (RDP), and clustered into Operational Taxonomic

Units (OTUs) using a 97% similarity cutoff with an average neighbor clustering algorithm. The number of sequences for each sample was rarefied to 10523 to minimize uneven sampling.

Lesion Model Creation: The Random Forest [19] algorithm was used to create the model used for prediction of lesion (adenoma or carcinoma) with the main training and testing of the model completed on an independent data set of 423 individuals. This model was then applied to our 67-person cohort. It should be noted that all individuals with an adenoma or carcinoma were grouped together to form the lesion group and the model was not created to find differences between normal, adenoma, and carcinoma but rather differences between both adenoma and carcinoma versus normal.

The model included only OTU data obtained from 16S rRNA sequencing. Non-binary data was checked for near zero variance and OTUs that had near zero variance were removed. This pre-processing was performed with the R package caret (v6.0.73). Optimization of the mtry hyper-parameter involved making 100 different 80/20 (train/test) splits of the data where normal and lesion were represented in the same proportion within both the whole data set and the 80/20 split. For each a 20 repeated 10-fold cross validation was performed on 80% component to optimize the mtry hyper-parameter by maximizing the AUC (Area Under the Curve of the Receiver Operator Characteristic). The resulting model was then tested on the hold out data obtained from the 20% component. Assessment of the most important OTUs to the model involved counting the number of times an OTU was present in the top 10% of mean decrease in accuracy (MDA) for each of the 100 different splits run. This was then followed with filtering of this list to variables that were only present in more than 50% of these 100 runs. The final collated list of variables was then run through the mtry optimization again. Once the ideal mtry was found the entire 423 sample set was used to create the final Random Forest model on which classifications on the 67-person cohort was completed.

The default cutoff of 0.5 was used as the threshold to classify individuals as positive or negative for lesion. The hyper-parameter, mtry, defines the number of variables to investigate at each split before a new division of the data was created with the Random Forest model.

Treatment Model Creation: We also investigated whether a model could be created that could identify pre- (initial) and post- (follow up) treatment samples. The main difference was that only the 67-person cohort was used at all stages of model building and classification. Other than this difference the creation of this model and optimization of the mtry hyper-parameter was completed using the same procedure as was used for the lesion model. Instead of classifying samples as positive or negative of lesion this model classified samples as positive or negative for being an initial sample prior to treatment.

Statistical Analysis: The R software package (v3.3.2) was used for all statistical analysis. Comparisons between bacterial community structure utilized PERMANOVA [20] in the vegan package (v2.4.1). Comparisons between probabilities as well as overall OTU differences between initial and follow up samples utilized a paired Wilcoxon ranked sum test. Where multiple comparison testing was appropriate, a Benjamini-Hochberg (BH) correction was applied [21] and a corrected P-value of less than 0.05 was considered significant. Unless otherwise stated the P-values reported are those that were BH corrected.

Analysis Overview: We first tested for any differences based on whether the individual had an adenoma or carcinoma. This was done by testing initial and follow up samples for differences in alpha and beta diversity, testing all OTUs, and investigating the relative abundance of genera from previously associated CRC bacteria (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, and *Porphyromonas*). Next, the lesion model was tested for accuracy in prediction and whether it reduced the positive probability of lesion in follow up samples. We then used the treatment model to assess whether it could classify

294 samples better than the lesion model and whether it could reduce the positive probability
295 of an initial sample in the follow up samples. Common OTUs were found for the two
296 different models used to assess which were important for both models. Finally, differences
297 between those receiving chemotherapy and radiation versus those who received neither
298 were tested.

299 ***Reproducible Methods:*** A detailed and reproducible description of how the data were
300 processed and analyzed can be found at [https://github.com/SchlossLab/Size_followUps_](https://github.com/SchlossLab/Size_followUps_2017)
301 2017. Raw sequences have been deposited into the NCBI Sequence Read Archive
302 (SRP062005 and SRP096978) and the necessary metadata can be found at [https://www.](https://www.ncbi.nlm.nih.gov/Traces/study/)
303 [ncbi.nlm.nih.gov/Traces/study/](https://www.ncbi.nlm.nih.gov/Traces/study/) and searching the respective SRA study accession.

Figure 1: General Differences between the Adenoma and Carcinoma Group. A) A significant difference was found between the adenoma and carcinoma group for thetacyc (P-value = 0.000472). Advanced adenomas are denoted as Screen Relevant Neoplasia (SRN). B) NMDS of the initial and follow up samples for the adenoma group. C) NMDS of the initial and follow up samples for the carcinoma group.

Figure 2: The Lesion Model. A) ROC curve: The shaded areas represent the range of values of a 100 different 80/20 splits of the test set data and the blue line represents the model using 100% of the data set and what was used for subsequent classification. B) Summary of Important Variables. MDA of the most important variables in the lesion model. The black point represents the mean and the different colors are the values of each different run up to 100. C) Positive probability change from initial to follow up sample in those with carcinoma. D) Positive probability change from initial to follow up sample if those with adenoma or advanced adenoma (Screen Relevant Neoplasia (SRN)).

Figure 3: The Treatment Model. A) ROC curve: The shaded areas represent the range of values of a 100 different 80/20 splits of the test set data and the blue line represents the model using 100% of the data set and what was used for subsequent classification. B) Summary of Important Variables. MDA of the most important variables in the initial sample model. The black point represents the mean and the different colors are the values of each different run up to 100. C) Positive probability change from initial to follow up sample in those with carcinoma. D) Positive probability change from initial to follow up sample of those with adenoma or advanced adenoma (Screen Relevant Neoplasia (SRN)).

Figure 4: Chemotherapy and Radiation Therapy Effects on Community. A) Treatment model initial sample positive probability reduction for chemotherapy. B) Treatment model initial sample positive probability reduction for radiation therapy. C) Common OTUs based on whether chemotherapy was received or not. The * denotes a P-value < 0.05 after multiple comparison correction. D) Common OTUs based on whether radiation therapy

330 was received or not.

Figure S1: Distribution of P-values from Paired Wilcoxon Analysis of All OTUs for Initial versus Follow Up

Figure S2: Previously Associated CRC Bacteria in Initial and Follow Up Samples.

A) Carcinoma initial and follow up samples had an observed significant difference in initial and follow up sample for the OTUs classified as *Parvimonas* (P-value = 0.0059), *Porphyromonas* (P-value = 0.225), *Peptostreptococcus* (P-value = 0.00424). B) Adenoma initial and follow up samples. There were no significant differences between initial and follow up (P-value = 0.881).

Declarations

Ethics approval and consent to participate

The University of Michigan Institutional Review Board approved this study, and all subjects provided informed consent. This study conformed to the guidelines of the Helsinki Declaration.

Consent for publication

Not applicable.

Availability of data and material

A detailed and reproducible description of how the data were processed and analyzed can be found at https://github.com/SchlossLab/Size_followUps_2017. Raw sequences have been deposited into the NCBI Sequence Read Archive (SRP062005 and SRP096978) and the necessary metadata can be found at <https://www.ncbi.nlm.nih.gov/Traces/study/> and searching the respective SRA study accession.

Competing Interests

All authors declare that they do not have any relevant competing interests to report.

Funding

This study was supported by funding from the National Institutes of Health to P. Schloss (R01GM099514, P30DK034933) and to the Early Detection Research Network (U01CA86400).

Authors' contributions

All authors were involved in the conception and design of the study. MAS analyzed the data. NTB processed samples and analyzed the data. All authors interpreted the data. MAS and PDS wrote the manuscript. All authors reviewed and revised the manuscript. All authors read and approved the final manuscript.

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 Amanda Elmore for reviewing and correcting code error and providing feedback on manuscript drafts. We would also like to thank Nicholas Lesniak for providing feedback on manuscript drafts.

References

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. *CA: a cancer journal for clinicians*. 2010;60:277–300.
2. Haggard FA, Boushey RP. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. *Clinics in Colon and Rectal Surgery*. 2009;22:191–7.
3. Zackular JP, Baxter NT, Chen GY, Schloss PD. Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. *mSphere*. 2016;1.
4. Arthur JC, Gharaibeh RZ, Mühlbauer M, Perez-Chanona E, Uronis JM, McCafferty J, et al. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. *Nature Communications*. 2014;5:4724.
5. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL, Rossetti BJ, et al. Microbiota organization is a distinct feature of proximal colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111:18321–6.
6. Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, et al. The gut microbiome modulates colon tumorigenesis. *mBio*. 2013;4:e00692–00613.
7. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*. 2016;8:37.
8. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*. 2014;10:766.
9. Flynn KJ, Baxter NT, Schloss PD. Metabolic and Community Synergy of Oral Bacteria in

Colorectal Cancer. *mSphere*. 2016;1.

10. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*. 2017;66:70–8.

11. Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research (Philadelphia, Pa.)*. 2014;7:1112–21.

12. Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, et al. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*. 2013;1:16.

13. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology* [Internet]. 2014 [cited 2017 Feb 14];12:661–72. Available from: <http://www.nature.com/doifinder/10.1038/nrmicro3344>

14. Zhu Y, Lin X, Li H, Li Y, Shi X, Zhao F, et al. Intake of Meat Proteins Substantially Increased the Relative Abundance of Genus *Lactobacillus* in Rat Feces. *PloS One*. 2016;11:e0152678.

15. Mu C, Yang Y, Luo Z, Guan L, Zhu W. The Colonic Microbiome and Epithelial Transcriptome Are Altered in Rats Fed a High-Protein Diet Compared with a Normal-Protein Diet. *The Journal of Nutrition*. 2016;146:474–83.

16. Ozdal T, Sela DA, Xiao J, Boyacioglu D, Chen F, Capanoglu E. The Reciprocal Interactions between Polyphenols and Gut Microbiota and Effects on Bioaccessibility. *Nutrients* [Internet]. 2016 [cited 2017 Feb 14];8:78. Available from: <http://www.mdpi.com/2072-6643/8/2/78>

17. O'Brien CL, Allison GE, Grimpen F, Pavli P. Impact of colonoscopy bowel preparation

414 on intestinal microbiota. PloS One. 2013;8:e62815.

415 18. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a
416 dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence
417 data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology.
418 2013;79:5112–20.

419 19. Breiman L. Random Forests. Machine Learning [Internet]. 2001 [cited 2013 Feb
420 7];45:5–32. Available from: <http://link.springer.com/article/10.1023/A%3A1010933404324>
421 <http://link.springer.com/article/10.1023%2FA%3A1010933404324?LI=true>

422 20. Anderson MJ, Walsh DCI. PERMANOVA, ANOSIM, and the Mantel test in the face of
423 heterogeneous dispersions: What null hypothesis are you testing? Ecological Monographs
424 [Internet]. 2013 [cited 2017 Jan 5];83:557–74. Available from: [http://doi.wiley.com/10.1890/](http://doi.wiley.com/10.1890/12-2010.1)
425 12-2010.1

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