Differences in the Stool Microbiome Before and After Colorectal Cancer Treatment

Running Title: Human Microbiome and Colorectal Cancer

Marc A Sze, Nielson T Baxter, Mack T Ruffin IV, Mary AM Rogers, and Patrick D Schloss†

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.

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

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

Abstract

Colorectal cancer (CRC) continues to be a worldwide health problem with early detection being used as a key component in mitigating deaths due to the disease. Previous research suggests that the bacterial microbiome can be used as a biomarker for colorectal cancer (CRC). These reports have mostly focused on investigating how the bacterial microbiome is used at a single point in time to predict disease. In this study, we assessed whether a model built with bacterial microbiome data could accurately predict adenoma or carcinoma and adjust positive probability for lesion (adenoma or carcinoma) after surgical treatment. This model was tested on a 66 person group that included samples before and after treatment to allow for the assessment of how the model adjusts risk after treatment. The model chosen had a 10-fold cross validated AUC of 0.819 and a test set AUC of 0.815. : For the follow up samples our Random Forest model significantly decreased the positive probability of lesion compared to the initial samples for both adenoma (P-value = 7.95e-08) and carcinoma (P-value = 7.95e-08). In the test set, the prediction of lesion for the initial samples had a sensitivity of 100% (66/66) while follow up it was 100% (1 / 1). Our model predicted that 36.4% of the 67-person cohort had normal colons and 63.6% had a lesion. Our model suggests that treatment does significantly reduce the probability of having a colonic lesion. Further surveillance of these individuals will enable us to determine whether models such as the one we present here can also be used to predict recurrence of colorectal cancer.

Importance

This is one of the first studies to investigate within humans what happens to the bacterial microbiome before and after colorectal cancer treatment. Specifically, it aims to assess how the random forest machine learning algorithm built models respond to treatment and adjust their positive probability calls of whether the individual has an adenoma or carcinoma.

Introduction

Colorectal cancer (CRC) continues to be a leading cause of cancer related deaths and is the second most common cancer death among men aged 40-79 years of age (**insert citation**). Over the last few years death due to the disease has seen a significant decrease thanks mainly to improvements in screening (**insert citation**). However, despite this giant improvement there are still approximately 50,000 deaths from the disease a year (**insert citation**). It is estimated that around 5-10% of all CRCs can be explained by autosomal dominant inheritance (**insert citation**). The vast majority of CRCs are not inherited and the exact etiology to disease is not well worked out (**insert citation**).

Recent reports have found that the microbiome between those with and without CRC are different and that bacteia associated with CRC tissue is different that that found on the normal mucosa (**Insert citations**). Further, there is evidence to suggest that more inflammatory mouth-associated bacteria are more prevalent in the guts of those with CRC (**insert citation**). This has led to the hypothesis that these bacteria replace traditional residents that produce metabolites such as butyrate (**insert citation**) and lead to a more inflammatory state (**insert citation**). This inflammatory state allows for more of these types of bacteria to populate the gut and leads to neoplasia and eventually CRC (**insert citation**).

Building upon these findings a number of groups, including our own, have investigated using the bacterial microbiome as a potential biomarker for adenoma and CRC detection (**insert citations**). These collectively show that a reasonable area under the curve (AUC) can be obtained when using the stool bacterial microbiome to classify disease state (**insert citation**). Further, a number of groups have extended this observation to show that using either specific bacterial species or the bacterial microbiome as a whole paired with the Fecal Immunochemical Test (FIT) can either increase the AUC of the model (**insert**

citations).

In this study we further refine the random forest model by including more background information on the indivdiuals within it. We use a rigourous and accepted approach for machine learning training and testing to validate this model (**insert citation**). We then apply it to individuals in which we have samples before and after treatment to assess how well the model adjusts their probability of having an adenoma or carcinoma. Finally, we investigate which specific Operational Taxonomic Units (OTUs) are most affected by treatment. Within this study lesion refers to both adenoma and carcinoma.

Results

Bacterial Community and Fit Changes before and after Treatment Based on thetayc distance metrics comparing the initial to the follow up samples there was no difference between the adenoma and carcinoma group (P-value = 0.697) [Figure 1a]. There was a difference in FIT between initial and follow up samples with the carcinoma group having a significant decrease in FIT versus the adenoma group (P-value = 2.15e-05) [Figure 1b]. Although the thetayc distance metric change was similar between adenoma and carcinoma the directionality of the change was significant in the cancer group between initial and follow up (P-value = 0.001) but not for the adenoma group (P-value = 1) [Figure 2]. When all follow up samples were compared to each other there was no significant overall difference between them (P-value = 0.643). There was no significant difference between initial and follow up samples for observed OTUs, Shannon diversity, and evenness after correction for multiple comparisons [Table S1]. Time of follow up sample from initial sampling did not have an effect on our prediction data set (uncorrected P-value = 0.784).

Outcome of Model Training The range of the AUC for model training ranged from a minimum of 0.795 to a maximum of 0.854. To be conservative the model chosen for prediction of disease in the follow up samples had an AUC in the middle of all the 100 runs which was 0.819. Interestingly, the worst AUC model from training performed the best on it's respective 80/20 split test data [Figure 3]. In fact the 80/20 test performance showed that the AUC for the middle model chosen was the most stable (best training model test set AUC = 0.682, middle training model test set AUC = 0.815, worse training model test set AUC = 0.931). That is to say it had the smallest change in AUC in comparison to the other minimum and maximum AUC trained models. There was no significant difference between the AUC of the best and middle training models (P-value = 0.419). There was also no difference in the middle model versus worse (P-value = 0.178) or full data model (P-value = 1). The two comparisons with a significant difference were between the worse

training model and the best training model (P-value = 7.94e-04) and the worse training model and full data model (P-value = 3.52e-03).

Most Important Variables to the Model Overall, there were a total of 32 variables identified as being present in more than 50% of the training models [Table S2]. The top 5 most important bacterial OTUs were Bacteria (Otu000013), Escherichia/Shigella (Otu000018), Bacteria (Otu000020), Ruminococcus (Otu000017), and Porphyromonas (Otu000153). These 5 OTUs were present in at least 99 out of the total 100 different 80/20 runs.

Surgical Removal of an Adenoma or Carcinoma Results in a Decrease in Positive Probability Prediction A total of 1 sample was omitted from the original 67 sample set since it was missing a complete set of follow up data. This left a total of 66 samples for test predictions. After multiple comparison correction there was a significant overall decrease in positive probability of a carcinoma and adenoma (P-value = 1.14e-11) [Figure 4]. This decrease was significant for both adenoma (P-value = 7.95e-08 [Figure 4a] and carcinoma (P-value = 7.95e-08) [Figure 4b] alone. This also held specifically for those with screen relative neoplasias (SRN) (P-value = 5.35e-04). A total of 66 or 100% of all samples were correctly predicted to have an adenoma or carcinoma. Although there was a decrease in positive probability only 24 of the total 66 individuals were classified as adenoma or carcinoma free on follow up (successful classification of 37.9%). There was no significant difference between the predictions and actual diagnosis for the initial samples in the 67-sample cohort test set. However, the predictions were significantly discordent with the diagnosis for the follow up samples (P-value = 4.19e-10). Although there were discordent results the respective sensitivity for the initial group was 100% and for follow up was 100%, respectively.

There was 1 individual who still clearly had CRC on follow up as well as 5 individuals whose status on follow up was unknown. Although the 1 individual had a decrease in positive

probability their follow up sample was still higher than the cutoff threshold of 0.5 (positive probability = 0.927). Interestingly, 2 individuals who were unknown on follow up still were over the threshold cutoff of 0.5 even though, like the 1 individual with clear CRC on follow up, the probability of an adenoma or carcinoma decreased [Table S3].

The follow up positive probabilities were not affected by either chemotherapy treatment (uncorrected P-value = 0.919) or radiation therapy (uncorrected P-value = 1). There was also no difference in the amount of change in the positive probability based on whether individuals received chemotherapy (unccorected P-value = 0.578) or radiation therapy (uncorrected P-value = 0.904).

Specific OTUs in the Lesion Model are not Detected in Follow Up Versus Initial Samples Overall, there were a total of 8 OTUs that were common between the main lesion model and the model for classifying initial and follow up samples specifically [Table S4]. A total of 1 OTU was still significant after multiple comparison correction and it's lowest taxonomic identification was to Blautia. In general, Otu000012 (Blautia) was decreased from initial to follow up [Figure 5]. The relative abundance was not drastically different then the mean of the values observed in the control training set [Figure 5].

Discussion

In our training set we show that the overall community structure as measured by different alpha diversity metrics, shows very little change between controls and those with either adenoma or carcinoma [Table S1]. With respect to our test set there was very little difference in magnitude of change in the thetayc distance metric between those with adenoma or carcinoma [Figure 1a]. In contrast, FIT had a large change in the initial and follow up samples in the carcinoma group versus the adenoma [Figure 1b]. An NMDS showed that there was very little observable change between initial and follow up for the adenoma group but there was one for the carcinoma group [Figure 2]. This cursory information is suggestive that treatment of carcinoma, had the largest response.

We next created a model that incorporated both patient metadata, FIT, and the bacterial microbiome to be able to predict lesions (adenoma or carcinoma). We chose the middle training model, based on AUC, from 100 80/20 (train/test) splits. It's 10-fold cross validated AUC was similar to it's test set AUC which was not the case for both the best and worse training model [Figure 3]. Using this model we predicted the probability of a lesion in the inital and follow up samples [Figure 4]. There was a signficant decrease in positive probability regardless of whether the sample as a carcinoma or adenoma. The overall sensitivity for lesion detection in the intial samples was `100` and for follow ups was `100`. Although there was a decrease in overall probability of an adenoma or carcinoma only `24` were below the 0.5 threshold out of the total `65` individuals who were diagnosed as not having a carcinoma on follow up.

We then investigated which OTUs could potentially be more important in our model [Figure 5 & Table S4]. Only a single OTU was significant after multiple comparison correction and the lowest taxonomic identification of `Otu000012` was to `Blautia`. Although there was a difference in the relative abundance at initial and follow up these values were not drastically

different from the relative abundance values observed in the control individuals of the training set [Figure 5]. This research provides evidence that it is possible to create a highily sensitive model for detection of adenoma or carcinoma with bacterial microbiome data. It accomplishes this by using a unique sample set in which before and after stool samples are available for assessment. By using these types of samples we are able to show that this model is reactive. That is to say that after surgery for removal of the adenoma or carcinoma it decreases the positive probability to reflect a lower likelihood of the individual having an adenoma or carcinoma.

This study builds upon previous work from numerous labs that have looked into the bacterial microbiome as a potential screening tool (**insert citation**). Based on previous work by Jobin, et al. (**insert citation**) it may not be surprising to see E.coli in the top 5 OTUs for the training models. Similarily, Porphyromonas has also been implicated in colorectal cancer (**insert citation**). Interestingly, the other OTUs had taxonomic identification for resident gut microbes. This could suggest that changes to the resident microbiome are important to the initiation of adenoma or carcinoma formation (**insert citation**) and provide support for the hypothesis that an initial change in the bacterial microbiome could pave the way for more inflammatory species: whether by creation of a new niche for oral microbes (**insert citation**) or allowing for a bloom of existing pro-inflammatory residents (**insert citation**).

Naturally, it is curious that normal staples of many screening studies such as Fusobacterium, Parvimonas, and Peptostreptococcus were not present in the majority of the training models. One potential explanation for this is that FIT provides the same information to the model as these three organisms and so the model uses FIT preferentially over them. This has been suggested to be the case in a previous study (**insert Baxter Study**). Regardless, our study does not argue against the importance of these bacterium in CRC initiation or pathogenesis but rather that the best model does not utilize these specific bacteria for prediction purposes. Another potential reason why we did not identify

the usual suspects is that they may not change much between our initial and follow up samples. That is to say that they are consistently present even after removal of the lesion but surgery.

One limitation is that we do not know whether these individuals eventually had a CRC remission. This information would help to strengthen the case for our Random Forest based model keeping a number of individuals above the cutoff threshold even though at follow up they were diagnosed as no longer having CRC or adenomas. Another limitation is that we do not know if adding modern tests such as the stool DNA test (**insert citation**) could help improve our overall AUC. Finally, although our training and test set are relatively large we still run the risk of overfitting or having a model that may not be immediately extrapolateable to 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.

By adding patient data such as age, BMI, etc. to the model and showing that it can successfully predict both carcinoma and adenoma our study provides further data that these patient factors in conjunction with the bacterial microbiome could potentially influence CRC and perhaps have a role in formation of adenomas. Further studies need to be carried out to verify our findings since not only are we dealing with stool, which could be very different than the communities present on the actual tissue, but also are dealing with correlations.

Despite these limitations we think that these findings significantly add to the existing scientific knowledge on CRC and the bacterial microbiome. The ability for machine learning algorithms to take bacterial microbiome data and successfully lower positive probability after either adenoma or carcinoma removal provides evidence that there are specific signatures associated with these lesions. It also shows that these algorithms successfully react to successful treatment regimens and may be able to one day diagnose

CRC with a high level of accuracy.

Methods

Study Design and Patient Sampling The sampling and design of the study was similar to that reported in Baxter, et al (**insert citation**). In brief, study exclusion involved those who had already undergone surgery, radiation, or chemotherapy, had colorectal cancer before a baseline stool sample could be obtained, had IBD, a known hereditary non-polyposis colorectal cacner, or Familial adenomatous polyposis. Samples used to build the model used for prediction were collected either prior to a colonoscopy or between 1 - 2 weeks after. The bacterial microbiome has been shown to nomralize within this time period (**insert citation**). Kept apart from this training set were a total of 67 individuals that not only had a sample as described previoulsy but also a follow up sample between 188 - 546 days after surgery and treatment had been completed. 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.

Fecal Immunochemical Test and 16S rRNA Gene Sequencing FIT was analyzed as previously published using both OC FIT-CHEK and OC-Auto Micro 80 automated system (Polymedco Inc.) (**insert citation**). 16S rRNA gene sequencing was completed as previously described by Kozich, et al. (**insert citation**). In brief, 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 control, adenoma, and carcinoma evenly represented on each run. Which of each group was randomly assigned to avoid biases based on sample collection location.

Sequence Processing The mothur software package (v1.37.5) was used to process the 16S rRNA gene sequences. This process has been previously described (**insert citations**). The general processing workflow using mothur is as follows: Paired-end reads

were first merged into contigs, quality filtered, aligned to the SILVA database, screening 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 numer of sequences for each sample was rarified to `10521` in an attempt to minimize uneven sampling.

Model Creation The Random Forest (**insert citation**) algorithm was used to create the model used for prediction of lesion (adenoma or carcinoma) for the `67` individuals with follow up samples. The model included data on FIT, the bacterial microbiome, sex, age, Body Mass Index (BMI), whether the indivdiual was caucasion or not, history of cancer, and family history of cancer. Non-binary data was checked for near zero variance and auto correlation. Data columns that had near zero variance were removed. Columns that were correlated with each other over a Spearman correlation coefficient of 0.75 had one of the two columns removed. This pre-processing was performed with the R package caret (`v6.0.73`). Optimiation of the mtry hyperparameter as well as data on the best and worse performance of the model involved taking the `490` samples and making 100 80/20 (train/test) splits in the data where control and lesion were equally represented in the 80 and 20 split, respectively. This 80% portion was then split again into an 80/20 split, and run twenty times through 10-fold cross validation to optimzie the model's Receiver Operator Characteristic (ROC) and tuneing both mtry and ntree. This resulting model was then tested on the 20% of the data that was originally held out from this overall process. This was repeated 100 times with the best model after all the repeats chosen to be used on the `67` samples with follow up. Once the ideal mtry was found the entire `490` sample set was used to create the final Random Forest model. The default cutoff of a probability of 0.5 was used as the threshold to classify individuals as positive or negative of lesion.

Selection of Important OTUs In order to assess which variables were most central to all the models we counted the number of times a variable was present in the top 10% of mean

decrease in accuracy (MDA) for each different 80/20 split model and then filtered this list to variables that were only present more than 50% of the time. This final collated list of variables was what was considered the most important.

A second model using the same types of conditions was used to identify OTUs that could predict initial and follow up samples. This new model OTUs were then compared to the overall lesion model and the resulting common OTUs to both models were selected to test for changes between initial and follow up using a paired wilcoxson rank sum test.

Statistical Analysis The R software package (`v3.3.0`) was used for all statisitical analysis. Comparisons between bacterial community structure utilized PERMANOVA (**insert citation**) in the vegan package (`v2.4.1`) while comparisons between ROC curves utilized the method by DeLong et al. (**insert citation**) executed by the pROC (`v1.8`) package. Comparisons between probabilities as well as overall amount of OTU between initial and follow up samples utilized a paired wilcoxson ranked sum test. Where multiple comparison testing was needed a Benjamini-Hochberg (BH) correction was applied (**insert citation**) and a corrected P-value of less than 0.05 was considered significant. Unless otherwise stated the P-values reported are those of the BH corrected ones.

Reproducible methods. A detailed and reproducible description of how the data were processed and analyzed can be found at https://github.com/SchlossLab/Baxter_followUps_2016.

Acknowledgements

The authors thank the Great Lakes-New England Early Detection Research Network for providing the fecal samples that were used in this study. 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).

Figure 1: Change in Thetayc and Fit between initial and follow up in adenoma or carcinoma group. A) No significant difference was found between the adenoma and carcinoma group for thetayc (P-value = 0.697). B) A significant difference was found between the adenoma and carcinoma group for Fit (P-value = 2.15e-05).

Figure 2: NMDS of the Overall Bacterial Community Changes. A) NMDS of the intial and follow up samples for the Adenoma group. B) NMDS of the initial and follow up samples for the Carcinoma group.

Figure 3: Graph of the Receiver Operating Characteristic Curve on Test Set Performance of the Best, Middle, and Worse Training Models. For each of the 100 training cohort sets used had 392 individuals and the testing cohort sets had 13 individuals. The AUC on the test sets for the best, middle, and worse models from training were 0.682, 0.815, and 0.931, respectively. cvAUC is the 10-fold cross-validated AUC from training.

Figure 4: Breakdown by Carcinoma and Adenoma of Prediction Results for Initial and Follow Up* A) Positive probability adjustment of those with carcinoma from intial to follow up sample B) Positive probability adjustment of those with adenoma as well as those with SRN and the probability adjustment from initial to follow up sample. THe dotted line represents the threshold used to make the decision of whether a sample was lesion positive or not.

Figure 5: Lesion Model OTU with a Significant Decrease in Relative Abundance that is also Predictive of Initial and Follow Up. After multiple comparison correction 1 (Blautia) was the only one with a P-value < 0.05. The dotted line represents the average relative abundance in the control training group.

Figure S1: Thetayc Graphed Against Time of Follow up Sample from Initial

References