Making Sense of the Noise: Leveraging Existing 16S rRNA Gene Surveys to Identify Key Community Members in Colorectal Tumors

Marc A Sze[1] 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

Co-author e-mails:

- marcsze@med.umich.edu

Abstract

Background. An increasing body of literature suggests that there is a crucial role for the microbiota in colorectal cancer (CRC) pathogenesis. Important drivers within this context have ranged from individual microbes to the whole community. Our study expands on a recent meta-analysis investigating microbial biomarkers for tumors by testing the hypothesis that the bacterial community has important associations to both adenoma and carcinoma tumors. To test this hypothesis we examined both feces (n = 1737) and colon tissue (492 total samples from 350 individuals) across 14 previously published 16S rRNA gene sequencing studies on colorectal tumors and the microbiota.

Results. Fecal samples had a significant decrease for both Shannon diversity and evenness as tumor severity increased, after correcting for study effect and variable region sequenced (P-value < 0.05). This reduction in evenness translated into small increases in the odds ratio for individuals to have both adenoma (P-value = 0.035) and carcinoma tumors (P-value = 0.0021) while the reduction in Shannon diversity only translated into an increased odds ratio for individuals to have carcinomas (P-value = 0.0069). Increases in mouth-associated microbes were commonly in the top 5 most increased odds ratios for individuals to have either adenoma or carcinoma tumors, regardless of sample type. Prediction models built to classify either individuals with adenoma or carcinoma were trained on the whole community or selected genera (top 5 highest and lowest odds ratios) from either fecal or tissue samples. Both the full and select models for either adenoma or carcinoma resulted in similar classification success according to Area Under the Curve (AUC). The most important groups within the full community models consistently belonged to genera such as *Ruminococcus*, *Bacteroides*, and *Roseburia* across studies. Although a number of associations between the microbiota and tumor were identified, the majority of studies that we used in this meta-analysis were only individually adequately powered for large effect sizes.

Conclusions. These data provide support for the importance of the bacterial community to both adenoma and carcinoma tumorgenesis. The evidence collected within this study on the role of the microbiota in those with tumors identifies a number of correlations that may not have been detected because of the low power associated with the majority of studies that have been performed to date.

Keywords

microbiota; colorectal cancer; polyps; adenoma; tumor; meta-analysis.

Background

Colorectal cancer (CRC) is a growing world-wide health problem in which the microbiota has been purported to play an active role in disease pathogenesis [1,2]. Numerous studies have shown the importance of both individual microbes [3–7] and the overall community [8–10] in tumorgenesis using mouse models of CRC. There have also been numerous case-control studies investigating the microbiota in the formation of both adenoma and carcinoma. A recent meta-analysis investigated whether specific biomarkers could be consistently identified using multiple data sets [11]. This meta-analysis focused on identifying microbial signatures of tumors (biomarkers) but did so on a small total number of individuals and only investigated stool. This present meta-analysis addresses some of these major shortcomings.

Although there has been an intense focus on microbiota-based biomarker discovery for tumors, the number of candidate genera seem to be endless. Some studies point towards mouth-associated genera such as *Fusobacterium*, *Peptostreptococcus*, *Parvimonas*, and *Porphyromonas* as key enriched genera [6,12–18]. Yet, even in these studies, mouth-associated genera are far from the only microbes identified to be associated with tumors. These other genera include, but are not limited to, *Providencia*, *Mogibacterium*, *Enterococcus*, *Escherichia/Shigella*, *Klebsiella*, and *Streptococcus* [14–16]. In fact, there is good *in vivo* evidence that *Escherichia/Shigella* and *Streptococcus* can be important in the pathogenesis of CRC [5,19,20]. Other studies have also identified *Akkermansia muciniphila* and *Bacteroides fragilis* as potential markers of CRC with good mechanistic studies for the latter [15,21,22]. A recent meta-analysis confirmed the correlations of certain mouth-associated genera and *Akkermansia muciniphila* with carcinoma [11]. However, the sample size (n = 509) is equal to or less than some of the more recent individual studies investigating the microbiota and colorectal tumors, making it hard to know how extrapolatable these findings are. That particular meta-analysis also

added more potential microbial associations to both carcinoma (*Pantoea agglomerans* *Ruminococcus*, *Lactobacillus*) and adenoma (*Prevotella*, *Methanosphaera*, *Succinovibrio*, *Haemophilus parainfluenzae*, *Ruminococcus*, *Lactobacillus*) stages of disease that need to be investigated further, since a number of these genera have been found to be enriched in controls and not disease [13,16,17]. Additionally, genera like *Roseburia* have been found in some studies to be increased in tumors but in others to either be decreased or have no difference [15,18,23,24].

Most of these studies have focused on individuals with carcinomas but associations with the adenoma stage of disease are not any clearer at identifying candidate genera correlated with these earlier tumors. Groups focusing on broad scale community metrics have found that measures such as richness are decreased in those with adenomas versus controls. Other studies have identified *Lactococcus*, *Pseudomonas*, *Acidovorax*, *Cloacibacterium*, *Helicobacter*, *Lactobacillus*, *Bilophila*, *Desulfovibrio*, and *Mogibacterium* to be increased in those with adenoma tumors [25–27]. Additionally, based on these studies mentioned, there seems to be very few common genera that are associated with both adenoma and carcinoma tumors, with *Lactobacillus* being one of the few commonalities.

Targeting the identification of tumor microbial biomarkers within stool seems logical since it offers an easy and cost-effective way to stratify risk of disease. The current gold standard for diagnosis, a colonoscopy, can be time-consuming and is not without risk of complications. Although stool represents an easy and less invasive way to assess risk, it is not clear how well this sample reflects adenoma- and carcinoma- associated microbial communities. Some studies have tried to assess this in health and disease but are limited by their sample size [18,28]. Sampling the microbiota directly associated with colon tissue may provide clearer answers but is not without their own limitations. After the colonoscopy bowel prep the bacterial community sampled may reflect the better adhered microbiota versus the resident community. Additionally, these samples contain more host

DNA, potentially limiting the types of analysis that can be done. It is well known that low biomass samples can be very difficult to work with and results can be study dependent due to the randomness of contamination [29].

In comparison to the previous meta-analysis, this study significantly increases the total stool samples investigated, re-examines important genera across adenoma and carcinoma across study, and examines differences and similarities between stool and tissue microbiota in the context of colorectal tumors. Importantly, this analysis and approach could provide valuable insights into the common genera that are both protective and detrimental in individuals with adenoma or carcinoma and whether broad bacterial community measurements can account for these changes that were not provided by earlier meta-analysis studies [11].

Using both feces (n = 1737) and colon tissues (492 samples from 350 individuals) totaling over 2229 total samples across 14 studies [12–18,21,23–27,30] [Table 1 & 2], we expand both the breadth and scope of the previous meta-analysis to investigate whether the bacterial community or specific members are more important risk factors for both adenoma and carcinoma stages of disease. To accomplish this we first assessed whether bacterial diversity changes throughout disease (control to adenoma to carcinoma) and if it results in an increased odds ratio (OR) for individuals to have either an adenoma or carcinoma. We then assessed what genera, if any, increase or decrease the OR of an individual to have an adenoma or carcinoma. Next, using Random Forest models, we analyzed whether the full community or only the combined top 5 increased and top 5 decreased OR genera resulted in better model classification, based on the area under the curve (AUC). Finally, we also examined at what effect and sample size the studies used were powered for and the sample size needed to get to the traditionally accepted 80% power. Our results from these analyses suggests that the bacterial community changes as disease severity worsens, that specific members are important for disease classification, and that many of

the individual studies are underpowered for assessing small effect sizes.

Results

Lower Bacterial Diversity is Associated with Increased OR of Tumors: To assess differences in broad scale community metrics as disease severity worsens Operational Taxonomic Unit (OTU) richness, evenness, and Shannon diversity measurements were power transformed and Z-score normalized. These metrics are commonly used to assess the total number of OTUs, the equality of their abundance, and the overall diversity, respectively. Using linear mixed-effect models to control for study and variable region we assessed whether OTU richness, evenness, or Shannon diversity changed in a step-wise manner with disease severity. In stool, there was a significant decrease in both evenness and Shannon diversity as disease severity moved from control to adenoma to carcinoma (P-value = 0.025 and 0.043, respectively) [Figure 1A]. We next tested whether the detectable differences in community significantly increased in OR of having an adenoma or carcinoma. For fecal samples, a decrease versus the overall median in evenness resulted in a significantly increased RR for carcinoma (OR = 1.66 (1.2 - 2.3), P-value = 0.0021) and adenoma (OR = 1.3 (1.02 - 1.65), P-value = 0.035) while a decrease versus the overall median in Shannon diversity only increased the OR for carcinoma (OR = 1.61 (1.14 - 2.28), P-value = 0.0069) [Figure 2]. Using the Bray-Curtis distance metric and PERMANOVA, it was also possible to identify significant bacterial community changes, in specific studies, for both carcinoma-associated and adenoma-associated microbiota versus control [Table S1 & S2].

Using similar transformations for tissue samples, linear mixed-effect models were used on the transformed combined data to control for study, re-sampling of the same individual, and 16S variable region to test whether OTU richness, evenness, or Shannon diversity changed in a step-wise manner as disease severity increased. For colon tissue, there were no significant changes in OTU richness, evenness, or Shannon diversity as disease severity progressed from control to adenoma to carcinoma (P-value > 0.05) [Figure 1B & C].

We next analyzed the RR, for matched (unaffected tissue and an adenoma or carcinoma from the same individual) and unmatched (control and adenoma or carcinoma tissue not from the same individual) colon tissue samples. For individuals at either an adenoma or carcinoma stage of disease there was no significant change in RR based on lower than median values for OTU richness, evenness, and Shannon diversity [Table S3-S5]. Similar to stool samples, significant differences in bacterial community, assessed by PERMANOVA, were identified in unmatched tissue samples, for those at either adenoma or carcinoma stage of CRC [Table S6 & S7]. For studies with matched samples no differences in bacterial community were observed when assessed with PERMANOVA [Table S6 & S7]. These tissue results suggest that the microbiota within an individual are similar to each other regardless of disease status.

Mouth-Associated Genera are Associated with an Increased OR of Tumor: Next, we asked if being higher than the median relative abundance, for any specific genera, resulted in an altered OR for adenoma or carcinoma, in stool and colon tissue, due to our previous observations of small increases in OR using OTU richness and Shannon diversity. To investigate this we analyzed all common genera across each study, in colon tissue or stool, and assessed whether a relative abundance higher than the median results in an increase or decrease in OR. Mouth-associated genera were commonly found in the top 5 genera associated with an increased OR of having an adenoma (*Porphyromonas* [Figure 3A] and *Rothia* [Figure 3C]) and carcinoma (*Fusobacterium*, *Parvimonas*, *Porphyromonas*, and *Peptostreptococcus* [Figure 3B] and *Fusobacterium* and *Parvimonas* [Figure 3D]) for both stool and colon tissue samples. Conversely, genera commonly associated with the gastrointestinal tract were correlated with a decreased OR for both adenoma and carcinoma for both stool and colon tissue samples [Figure 3]. Even though mouth-associated genera were identified across disease stage, there was little direct overlap of the top 5 increased or decreased OR genera between both stages and sample site.

When observing ORs for adenoma between genera from stool or colon tissue with a P-value less than 0.05 there was almost no overlap and when they were similar the OR was in opposite directions (e.g. *Lactococcus*) [Table S8 & S9]. Many of the adenoma associated genera ORs with a P-value under 0.05 for colon tissue are also highly prevalent in contamination, specifically, *Novosphingobium*, *Pseudomonas*, and *Achromobacter* [Figure 3 & Table S8-S9]. For carcinoma stage of disease, certain mouth-associated genera (*Fusobacterium*, *Parvimonas*) had an increased OR for both colon tissue and stool samples [Table S10 & S11]. The genera with the highest increased OR for carcinoma in tissue was *Lepttrichia* while in stool it was *Peptostreptococcus* [Table S10 & S11].

Select Community Models can Recapitulate Whole Community Models: Since specific genera increased the OR for carcinoma over diversity metrics we assessed whether the bacterial community was better at classifying disease versus only a select group of genera. We selected these genera based on their OR and P-value significance and used two approaches to test this question. The first approach used genus level data and tested for differences in AUC between all genera and selected genera. A single study was used for training the model prior to testing on all other studies and this was repeated for every study in the meta-analysis. The second approach used OTU level data and tested for a generalized decrease in the 10-fold cross validation (CV) model AUC which is a common approach used to guard against over-fitting. This was applied across study and the AUC of the all OTUs model was compared against the model that used only OTUs that taxonomically classified to selected genera.

For the first approach using the genera-based models, the training set median AUC for model classification was similar for both the full and select genera models, for both tissue and stool studies [Figure S2-S3]. When analyzing the tests sets that were comprised of genera data from other studies, both models had a similar ability to detect individuals with adenomas or carcinomas, with the select genera models performing better in some

instances [Figure S4-S6]. Conversely, the second approach that used OTU-based models showed a slight decrease in median AUC between the full and select models [Figure 4 & 5].

In stool, the most common genera in the top 10 most important variables, in the full community models using the first approach, were *Ruminococcus*, *Bacteroides*, and *Roseburia* [Figure 6A & B]. Regardless of sample type, mouth-associated genera were present in models for the carcinoma stage of CRC [Figure 6A & B]. Yet, none were present in the majority of studies and *Fusobacterium* was the only genus present in the adenoma stage of CRC [Figure 6A & B]. For the second approach that utilized full OTU-based models, *Ruminococcaceae* was present in the top 10 consistently for both adenoma and carcinoma models while *Roseburia* was only present in many adenoma models and *Bacteroides* was present in the overwhelming majority of the carcinoma models [Figure 6C & 6D].

Unlike the stool-based Random Forest models, the tissue-based models, for the full genera from the first approach, showed no consistent representation of *Ruminococcaceae*, *Ruminococcus*, *Bacteroides*, and *Roseburia* in the top 10 most important model variables across study [Figure S7]. The vast majority of the top 10 model variables for the genera- and OTU-based models using colon tissue tended to be study specific. Further, there was very little overlap in the top 10 important variables between adenoma and carcinoma stage models, regardless of whether colon tissue or stool was used [Figure 6 & S7]. This discordance between stool and colon tissue samples also applies to the mouth-associated genera with one noticeable skew being that *Fusobacterium* and *Fusobacteriaceae* occur more often in the top 10 of matched versus unmatched colon tissue Random Forest models [Figure S7B-C & S7E-F]. This suggests that either the colon tissue microbiota is study and person dependent or that kit and/or other types of contamination associated with low biomass samples may be skewing the results.

CRC Studies are Underpowered for Detecting Small Effect Sizes: Next, we assessed

how much confidence should be placed in the reported outcomes from each individual study by calculating the ability to detect a difference (power) and sample size needed for small, medium, and large effect size differences between cases and controls. When assessing the power of each study at different effect sizes the majority of studies achieved 80% power to detect a 30% or greater difference between groups [Figure 7A & B]. No study that we analyzed had the standard 80% power to detect an effect size difference equal to or below 10% [Figure 7A & B]. In order to achieve a power of 80%, for small effect sizes, studies used in our meta-analysis would need to recruit over 1000 individuals for both the case and control arms [Figure 7C]

Discussion

Our study identifies clear differences in diversity, both at the community level and for individual genera, present in patients with and without CRC [Figure 1-3]. Although there was a step-wise decrease in diversity as disease progressed from control to adenoma to carcinoma, this did not translate into large effect sizes for the OR of tumors. Even though mouth-associated genera increased the OR of having a carcinoma, they did not consistently increase the RR of having an adenoma. Additionally, our observations suggest that by combining mouth-associated and CRC protective microbes we can classify either adenoma or carcinoma stage of disease as well as models that use the full community.

The data presented herein support the importance of select genera for carcinoma, but not necessarily adenoma, formation. The results that we have presented show that both the genera and OTU select and full models, for the carcinoma stage of CRC, had similar AUCs [Figure 4 & 5]. This suggests that an interplay between a select number of potentially protective and exacerbating microbes within the GI community is crucial for carcinoma formation. Importantly, it suggests that there may be key members of the GI community that might be studied further to potentially reduce the risk of carcinoma. Conversely, using the present data, it is clear that new approaches may be needed to identify members of the community associated with adenoma stage of disease. Regardless of sample type and whether a full or select model was used, our Random Forest models consistently performed poorly. Yet, the step-wise decrease in diversity suggests that the adenoma-associated community is not normal but has changed subtly [Figure 1]. This change in diversity, at this early stage of disease, could be focal to the adenoma itself. One possible explanation is that how the host interacts with these subtle changes at early stages of the disease is what leads to a thoroughly dysfunctional community that is supportive of CRC genesis.

Within stool, common GI microbes were most consistently present in the top 10

genera or OTUs across studies [Figure 6]. Changes in *Bacteroides*, *Ruminococcaceae*, *Ruminococcus*, and *Roseburia* were consistently found to be in the top 10 most important variables across the different studies for both adenoma and carcinoma [Figure 6]. These data suggest that whether the non-resident bacterium is *Fusobacteria* or *Peptostreptococcus* may not be as important as how these bacteria interact with the changing resident community. Based on these observations, it is possible to hypothesize that small changes in community structure lead to new niches in which any one of the mouth-associated genera can gain a foothold, exacerbating the initial changes in community and facilitating the transition from adenoma to carcinoma stage of disease.

The colon tissue-based studies did not provide a clearer understanding of how the microbiota may be associated with tumors. Generally, the full OTU-based models of unmatched and matched colon tissue samples were concordant with stool samples showing that GI resident microbes were the most prevalent in the top 10 most important variables across study [Figure S7E & F]. Unlike in stool, *Fusobacterium* was the only mouth-associated bacteria consistently present in the top 10 most important variables of the full carcinoma stage models [Figure S7B-C & E-F]. The majority of the colon tissue-based results seem to be study specific with many of the top 10 taxa being present only in a single study. Additionally, the presence of genera associated with contamination, within the top 10 most important variables for the genera and OTU models is worrying. The low bacterial biomass of tissue samples coupled with potential contamination could explain why these results seem to be more sporadic than the stool results.

One important caveat to this study is that even though genera associated with certain species such as *Bacteroides fragilis* and *Streptococcus gallolyticus* subsp. *gallolyticus* were not identified, it does not necessarily mean that these specific species are not important in human CRC [20,22]. Since we are limited in our aggregation of the data to the genus level, it is not possible to clearly delineate which species are contributing to overall

disease progression. Our observations are not inconsistent with the previous literature on either *Bacteroides fragilis* or *Streptococcus gallolyticus* subsp. *gallolyticus*. As an example, the stool-based full community models consistently identified the genus *Bacteroides*, as well as OTUs that classified as *Bacteroides*, to be important model components across studies. This suggests that even though *Bacteroides* may not increase the RR of CRC and may not vary in relative abundance, like *Fusobacterium*, it is still important in CRC. Additionally, *Streptococcus gallolyticus* subsp. *gallolyticus* is a mouth-associated microbe, and the results from this study suggest that regardless of sample type, mouth-associated genera are commonly associated with an increased RR for both adenoma and carcinoma stage of disease.

The associations between the microbiota and adenoma stage of disease are inconclusive, in part, because many studies may not be powered effectively to observe small effect sizes. None of the studies analyzed were properly powered to detect a 10% or lower change between cases and controls. The results within our meta-analysis suggest that a small effect size may well be the scope in which differences consistently occur between controls and adenoma stage of disease. Future studies investigating adenoma stage and the microbiota need to take power into consideration to reproducibly study whether the microbiota contributes to polyp formation. In contrast to adenoma stage of disease, our observations suggest that most studies analyzed have sufficient power to detect many changes in the carcinoma-associated microbiota because of large effect size differences between cases and controls [Figure 7].

Conclusion

By aggregating together a large collection of studies analyzing both fecal and colon tissue samples, we are able to provide evidence supporting the importance of the bacterial community in colorectal tumors. Further, the data presented here suggests that mouth-associated microbes can gain a foothold within the colon and are commonly associated with the greatest OR of individuals having a carcinoma. No conclusive signal with these mouth-associated microbes could be detected for adenoma stage of disease. Our observations also highlight the importance of power and sample number considerations when investigating the microbiota and adenoma stage of disease due to possible subtle changes in the community. Overall, associations between the microbiota and individuals with carcinoma stage of disease were much stronger than with to those with adenoma stage of disease.

Methods

Obtaining Data Sets: The studies used for this meta-analysis were identified through the review articles written by Keku, *et al.* and Vogtmann, *et al.* [31,32] and additional studies not mentioned in the reviews were obtained based on the authors' knowledge of the literature. Studies that used tissue or feces as their sample source for 454 or Illumina 16S rRNA gene sequencing analysis and had data sets with sequences available for analysis were included. Some studies were excluded because they did not have publicly available sequences or did not have metadata in which the authors were able to share. After these filtering steps, the following studies remained: Ahn, *et al.* [12], Baxter, *et al.* [13], Brim, *et al.* [30], Burns, *et al.* [16], Chen, *et al.* [14], Dejea, *et al.* [24], Flemer, *et al.* [18], Geng, *et al.* [23], Hale, *et al.* [27], Kostic, *et al.* [33], Lu, *et al.* [26], Sanapareddy, *et al.* [25], Wang, *et al.* [15], Weir, *et al.* [21], and Zeller, *et al.* [17]. The Zackular [34] study was not included because the 90 individuals analyzed within the study are contained within the larger Baxter study [13]. After sequence processing, all the case samples for the Kostic study had 100 or less sequences remaining and was excluded, leaving a total of 14 studies that analysis could be completed on.

Data Set Breakdown: In total, there were seven studies with only fecal samples (Ahn, Baxter, Brim, Hale, Wang, Weir, and Zeller), five studies with only tissue samples (Burns, Dejea, Geng, Lu, Sanapareddy), and two studies with both fecal and tissue samples (Chen and Flemer). The total number of individuals analyzed after sequence processing for feces was 1737 [Table 1]. The total number of matched and unmatched tissue samples that were analyzed after sequence processing was 492 [Table 2].

Sequence Processing: For the majority of studies, raw sequences were downloaded from the Sequence Read Archive (SRA) (ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/) and metadata were obtained by searching the respective accession

number of the study at the following website: http://www.ncbi.nlm.nih.gov/Traces/study/. Of the studies that did not have sequences and metadata on the SRA, data was obtained from DBGap (n = 1, [12]) and directly from the authors (n = 4, [18,21,25,27]). Each study was processed using the mothur (v1.39.3) software program [35] and quality filtering utilized the default methods for both 454 and Illumina based sequencing. If it was not possible to use the defaults, the stated quality cut-offs, from the study itself, were used instead. Sequences that were made up of an artificial combination of two or more different sequences and commonly known as chimeras were identified and removed using VSEARCH [36] before *de novo* OTU clustering at 97% similarity was completed using the OptiClust algorithm [37].

Statistical Analysis: All statistical analysis after sequence processing utilized the R (v3.4.3) software package [38]. For OTU richness, evenness, and Shannon diversity analysis, values were power transformed using the rcompanion (v1.11.1) package [39] and then Z-score normalized using the car (v2.1.6) package [40]. Testing for α-diversity differences utilized linear mixed-effect models created using the lme4 (v1.1.15) package [41] to correct for study, repeat sampling of individuals (tissue only), and 16S hyper-variable region used. Relative risk was analyzed using both the epiR (v0.9.93) and metafor (v2.0.0) packages [42,43] by assessing how many individuals with and without disease were above and below the overall median value within each specific study. Relative risk significance testing utilized the chi-squared test. β-diversity differences utilized a Bray-Curtis distance matrix and PERMANOVA executed with the vegan (v2.4.5) package [44]. Random Forest models were built using both the caret (v6.0.78) and randomForest (v4.6.12) packages [45,46]. Power analysis and estimations were made using the pwr (v1.2.1) and statmod (v1.4.30) packages [47,48]. All figures were created using both ggplot2 (v2.2.1) and gridExtra (v2.3) packages [49,50].

Study Analysis Overview: OTU richness, evenness, and Shannon diversity was first

assessed for differences between controls, adenoma stage, and carcinoma stage using both linear mixed-effect models and OR. For each individual study the Bray-Curtis index was used to assess differences between control-adenoma and control-carcinoma. Next, all common genera were assessed for differences in OR for having an adenoma or carcinoma and ranked based on P-value. We then built Random Forest models based on the full or selected community (the top 5 increased and top 5 decreased OR based on P-value). Comparison between the full and selected models took two different approaches. In the first approach, models were trained on one study then tested on the remaining studies. This process was repeated for every study in the meta-analysis. In the second approach models were built using OTU level data and a 10-fold CV over 100 different iterations, based on random 80/20 splitting of the data, was used to generate a range of expected AUCs. For these OTU-based models, the selected model included all OTUs that had a taxonomic classification to a taxa in the top 5 increased and top 5 decreased OR based on P-value. Finally, the power of each study was assessed for an effect size ranging from 1% to 30% and an estimated sample size, for these effect sizes, was generated based on 80% power. For comparisons in which only control versus adenoma individuals were made, the carcinoma individuals were excluded from each respective study. Similarly, for comparisons in which control versus carcinoma individuals were made the adenoma individuals were excluded from each respective study. For all analysis completed feces and tissue samples were kept separate. Within the tissue groups the data were further divided between samples from the same individual (matched) and those from different individuals (unmatched).

Obtaining Genera Relative Abundance and Selected Models: For the genera analysis of the OR, OTUs were added together based on the genus or lowest available taxonomic classification level and the total average counts, for 100 different subsamplings. The select models for the first approach utilized the top 5 most significantly increased and decreased ORs. For this approach, the full community models for each study were built

by utilizing all genera and lowest taxonomic groups identified within that particular study. The select models for the second approach, that used OTU-based Random Forest models, utilized a similar method as the first approach. The main difference was that any OTU that taxonomically classified to one of the genera in the top 5 increased or decreased OR were included in the select model. OTU Random Forest models using the full community included all OTUs.

Matched versus Unmatched Tissue Samples: In general, tissue samples with control and tumor samples from different individuals were classified as unmatched while samples that belonged to the same individual were classified as matched. Studies with matched data included Burns, Dejea, Geng, and Lu while those with unmatched data were from Burns, Flemer, Chen, and Sanapareddy. For some studies samples became unmatched when a corresponding matched sample did not make it through sequence processing. All samples, from both tissue sample types, were analyzed together for the linear mixed-effect models with samples from the same individual corrected for. For all other analysis, not mentioned herein, matched and unmatched samples were analyzed separately using the statistical approaches mentioned in the Statistical Analysis section.

Assessing Important Random Forest Model Variables: Using Mean Decrease in Accuracy (MDA) the top 10 most important variables to the Random Forest model were obtained for the full models of the two different approaches used. For the first approach utilizing genus-based models, the number of times that a genus showed up in the top 10 of the training set across each study was counted. For the second approach, that utilized the OTU-based models, the medians for each OTU across 100 different 80/20 splits of the data was generated and the top 10 OTUs then counted for each study. Common taxa, for the OTU based models, were identified by using the lowest classification within the RDP database for each of the specific OTUs obtained from the previous counts and the number of times this classification occurred in the top 10, in each study, was recorded. Finally, for

the two studies that had adenoma tissue (Lu and Flemer) were equally divided between matched and unmatched studies and were grouped together for the counting of the top 10 genera and OTUs.

Reproducible Methods: The code and analysis can be found at https://github.com/SchlossLab/Sze_CRCMetaAnalysis_Microbiome_2017. Unless otherwise mentioned, the accession number of raw sequences from the studies used in this analysis can be found directly in the respective batch file in the GitHub repository or in the original manuscript.

Declarations

Ethics approval and consent to participate

Ethics approval and informed consent for each of the studies used is mentioned in the respective manuscripts used in this meta-analysis.

Consent for publication

Not applicable.

Availability of data and material

A detailed and reproducible description of how the data were processed and analyzed for each study can be found at https://github.com/SchlossLab/Sze_CRCMetaAnalysis_Microbiome_2017. Raw sequences can be downloaded from the SRA in most cases and can be found in the respective study batch file in the GitHub repository or within the original publication. For instances when sequences are not publicly available, they may be accessed by contacting the corresponding authors from whence the data came.

Competing Interests

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

Funding

MAS is supported by a Canadian Institute of Health Research fellowship and a University of Michigan Postdoctoral Translational Scholar Program grant.

Authors' contributions

All authors helped to design and conceptualize the study. MAS identified and analyzed the data. MAS and PDS interpreted the data. MAS wrote the first draft of the manuscript and both he and PDS reviewed and revised updated versions. All authors approved the final manuscript.

Acknowledgements

The authors would like to thank all the study participants who were a part of each of the individual studies utilized. We would also like to thank each of the study authors for making their data available for use. Finally, we would like to thank the members of the Schloss lab for valuable feed back and proof reading during the formulation of this manuscript.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA: a cancer journal for clinicians. 2016;66:7–30.

2. Flynn KJ, Baxter NT, Schloss PD. Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. mSphere. 2016;1.

3. Goodwin AC, Destefano Shields CE, Wu S, Huso DL, Wu X, Murray-Stewart TR, et al. Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:15354–9.

4. Abed J, Emgård JEM, Zamir G, Faroja M, Almogy G, Grenov A, et al. Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. Cell Host & Microbe. 2016;20:215–25.

5. Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan T-J, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science (New York, NY). 2012;338:120–3.

6. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host & Microbe. 2013;14:207–15.

7. Wu S, Rhee K-J, Albesiano E, Rabizadeh S, Wu X, Yen H-R, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nature Medicine. 2009;15:1016–22.

8. Zackular JP, Baxter NT, Chen GY, Schloss PD. Manipulation of the Gut Microbiota

Reveals Role in Colon Tumorigenesis. mSphere. 2016;1.

9. 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.

10. Baxter NT, Zackular JP, Chen GY, Schloss PD. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. Microbiome. 2014;2:20.

11. Shah MS, DeSantis TZ, Weinmaier T, McMurdie PJ, Cope JL, Altrichter A, et al. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. Gut. 2017;

12. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, et al. Human gut microbiome and risk for colorectal cancer. Journal of the National Cancer Institute. 2013;105:1907–11.

13. 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.

14. Chen W, Liu F, Ling Z, Tong X, Xiang C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PloS One. 2012;7:e39743.

15. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. The ISME journal. 2012;6:320–9.

16. Burns MB, Lynch J, Starr TK, Knights D, Blekhman R. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome Medicine. 2015;7:55.

17. 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.

493 2014;10:766.

494 18. Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, et al.
495 Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut.
496 2017;66:633–43.

497 19. Arthur JC, Gharaibeh RZ, Mühlbauer M, Perez-Chanona E, Uronis JM, McCafferty J, et
498 al. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced
499 colorectal cancer. Nature Communications [Internet]. Springer Nature; 2014;5:4724.
500 Available from: https://doi.org/10.1038/ncomms5724

501 20. Aymeric L, Donnadieu F, Mulet C, Merle L du, Nigro G, Saffarian A, et al. Colorectal
502 cancer specific conditions promoteStreptococcus gallolyticusgut colonization. Proceedings
503 of the National Academy of Sciences [Internet]. Proceedings of the National Academy of
504 Sciences; 2017;115:E283–91. Available from: https://doi.org/10.1073/pnas.1715112115

505 21. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. Stool microbiome
506 and metabolome differences between colorectal cancer patients and healthy adults. PloS
507 One. 2013;8:e70803.

508 22. Boleij A, Hechenbleikner EM, Goodwin AC, Badani R, Stein EM, Lazarev MG, et
509 al. The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal
510 cancer patients. Clinical Infectious Diseases [Internet]. Oxford University Press (OUP);
511 2014;60:208–15. Available from: https://doi.org/10.1093/cid/ciu787

512 23. Geng J, Fan H, Tang X, Zhai H, Zhang Z. Diversified pattern of the human colorectal
513 cancer microbiome. Gut Pathogens. 2013;5:2.

514 24. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL, Rossetti BJ, et al.
515 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.

25. Sanapareddy N, Legge RM, Jovov B, McCoy A, Burcal L, Araujo-Perez F, et al. Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans. The ISME journal. 2012;6:1858–68.

26. Lu Y, Chen J, Zheng J, Hu G, Wang J, Huang C, et al. Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas. Scientific Reports. 2016;6:26337.

27. Hale VL, Chen J, Johnson S, Harrington SC, Yab TC, Smyrk TC, et al. Shifts in the Fecal Microbiota Associated with Adenomatous Polyps. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology. 2017;26:85–94.

28. Flynn KJ, Ruffin MT, Turgeon DK, Schloss PD. Spatial variation of the native colon microbiota in healthy adults. Cold Spring Harbor Laboratory; 2017; Available from: https://doi.org/10.1101/189886

29. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology [Internet]. Springer Nature; 2014;12. Available from: https://doi.org/10.1186/s12915-014-0087-z

30. Brim H, Yooseph S, Zoetendal EG, Lee E, Torralbo M, Laiyemo AO, et al. Microbiome analysis of stool samples from African Americans with colon polyps. PloS One. 2013;8:e81352.

31. Keku TO, Dulal S, Deveaux A, Jovov B, Han X. The gastrointestinal microbiota and colorectal cancer. American Journal of Physiology - Gastrointestinal and Liver Physiology [Internet]. 2015 [cited 2017 Oct 30];308:G351–63. Available from: http://ajpgi.physiology.

org/lookup/doi/10.1152/ajpgi.00360.2012

32. Vogtmann E, Goedert JJ. Epidemiologic studies of the human microbiome and cancer. British Journal of Cancer [Internet]. 2016 [cited 2017 Oct 30];114:237–42. Available from: http://www.nature.com/doifinder/10.1038/bjc.2015.465

33. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Research. 2012;22:292–8.

34. 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.

35. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. ApplEnvironMicrobiol [Internet]. 2009 [cited 12AD Jan 1];75:7537–41. Available from: http://aem.asm.org/cgi/content/abstract/75/23/7537

36. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

37. Westcott SL, Schloss PD. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere. 2017;2.

38. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/

39. Mangiafico S. Rcompanion: Functions to support extension education program

562 evaluation [Internet]. 2017. Available from: https://CRAN.R-project.org/package=
563 rcompanion

564 40. Fox J, Weisberg S. An R companion to applied regression [Internet]. Second. Thousand
565 Oaks CA: Sage; 2011. Available from: http://socserv.socsci.mcmaster.ca/jfox/Books/
566 Companion

567 41. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4.
568 Journal of Statistical Software. 2015;67:1–48.

569 42. Telmo Nunes MS with contributions from, Heuer C, Marshall J, Sanchez J, Thornton
570 R, Reiczigel J, et al. EpiR: Tools for the analysis of epidemiological data [Internet]. 2017.
571 Available from: https://CRAN.R-project.org/package=epiR

572 43. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of
573 Statistical Software [Internet]. 2010;36:1–48. Available from: http://www.jstatsoft.org/v36/
574 i03/

575 44. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Vegan:
576 Community ecology package [Internet]. 2017. Available from: https://CRAN.R-project.org/
577 package=vegan

578 45. Jed Wing MKC from, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T,
579 et al. Caret: Classification and regression training [Internet]. 2017. Available from:
580 https://CRAN.R-project.org/package=caret

581 46. Liaw A, Wiener M. Classification and regression by randomForest. R News [Internet].
582 2002;2:18–22. Available from: http://CRAN.R-project.org/doc/Rnews/

583 47. Champely S. Pwr: Basic functions for power analysis [Internet]. 2017. Available from:

584 https://CRAN.R-project.org/package=pwr

585 48. Giner G, Smyth GK. Statmod: Probability calculations for the inverse gaussian
586 distribution. R Journal. 2016;8:339–51.

587 49. Wickham H. Ggplot2: Elegant graphics for data analysis [Internet]. Springer-Verlag
588 New York; 2009. Available from: http://ggplot2.org

589 50. Auguie B. GridExtra: Miscellaneous functions for “grid” graphics [Internet]. 2017.
590 Available from: https://CRAN.R-project.org/package=gridExtra

Table 1: Total Individuals in each Study Included in the Stool Analysis

Study	Data Stored	16S Region	Control (n)	Adenoma (n)	Carcinoma (n)
Ahn	DBGap	V3-4	148	0	62
Baxter	SRA	V4	172	198	120
Brim	SRA	V1-3	6	6	0
Flemer	Author	V3-4	37	0	43
Hale	Author	V3-5	473	214	17
Wang	SRA	V3	56	0	46
Weir	Author	V4	4	0	7
Zeller	SRA	V4	50	37	41

Table 2: Studies with Tissue Samples Included in the Analysis

Study	Data Stored	16S Region	Control (n)	Adenoma (n)	Carcinoma (n)
Burns	SRA	V5-6	18	0	16
Chen	SRA	V1-V3	9	0	9
Dejea	SRA	V3-5	31	0	32
Flemer	Author	V3-4	103	37	94
Geng	SRA	V1-2	16	0	16
Lu	SRA	V3-4	20	20	0
Sanapareddy	Author	V1-2	38	0	33

Figure 1: Community Differences between Control, Adenoma, and Carcinoma Across Sampling Site. A) Stool sample community differences by disease group. B) Unmatched tissue samples differences by disease group. C) Matched tissue sample differences by group disease group. The dashed line represents a Z-score of 0 or no difference from the median.

Figure 2: Odds Ratio for Adenoma or Carcinoma based on Bacterial Community Metrics in Stool. A) Community-based odds ratio for adenoma. B) Community-based odds ratio for carcinoma. Colors represent the different variable regions used within the respective study.

Figure 3: Top 5 Genera that Decrease and Increase Odds Ratio for Lesion. A) Adenoma odds ratio in stool. B) Carcinoma odds ratio in stool. C) Adenoma odds ratio in tissue. D) Carcinoma odds ratio in tissue. For all panels the odds ratio was also compared to whether one, two, three, or four of the CRC-associated genera were present. Points represented as only half on the graph have an OR of infinity in the positive or negative direction.

Figure 4: Stool OTU Random Forest Model Across Studies. A) Adenoma random forest model between the full and select community OTUs only. B) Carcinoma random forest model between the full and select community OTUs only. The dotted line represents an AUC of 0.5 and the lines represent the range in which the AUC for the 100 different 80/20 runs fell between. The solid red line represents the median AUC of all the studies for either the full or select community OTUS only model.

Figure 5: Tissue OTU Random Forest Model Across Studies. A) Adenoma random forest model between the full and select community OTUs only. B) Carcinoma random forest model between the full and select community OTUs only. The dotted line represents an AUC of 0.5 and the lines represent the range in which the AUC for the 100 different

80/20 runs fell between. The solid red line represents the median AUC of all the studies for either the full community or select genera OTUS only model.

Figure 6: Most Common Genera Across Full Community Stool Study Models. A) Common genera in the top 10 for adenoma Random Forest genus models. B) Common genera in the top 10 for carcinoma Random Forest genus models. C) Common genera in the top 10 for adenoma Random Forest OTU models. D) Common genera in the top 10 for carcinoma Random Forest OTU models.

Figure 7: Power and Effect Size Analysis of Studies Included. A) Power based on effect size for studies with adenoma individuals. B) Power based on effect size for studies with carcinoma individuals. C) The estimated sample number needed for each arm of each study to detect an effect size of 1-30%. The dotted red lines in A) and B) represent a power of 0.8.

Figure S1: Odds Ratio for Adenoma or Carcinoma based on Bacterial Community Metrics in Tissue. A) Community-based odds ratio for adenoma. B) Community-based odds ratio for carcinoma. Colors represent the different variable regions used within the respective study.

Figure S2: Stool Random Forest Genus Model AUC for each Study. A) AUC of adenoma models using all genera or select genera only. B) AUC of carcinoma models using all genera or select genera only. The black line represents the median within each group.

Figure S3: Tissue Random Forest Genus Model AUC for each Study. A) AUC of adenoma models using all genera or only select genera divided between matched and unmatched tissue. B) AUC of carcinoma models using all genera or select genera only. The black line represents the median within each group divided between matched and unmatched tissue.

Figure S4: Stool Random Forest Prediction Success Using Genera Across Studies. A) AUC for prediction in adenoma using all genera or select genera only. B) AUC for prediction in carcinoma using all genera or select genera only. The dotted line represents an AUC of 0.5. The x-axis is the data set in which the model was initially trained on. The red lines represent the median AUC using that specific study as the training set.

Figure S5: Tissue Random Forest Prediction Success of Carcinoma Using Genera Across Studies. A) AUC for prediction in unmatched tissue for all genera or select genera only. B) AUC for prediction in matched tissue using all genera or select genera only. The dotted line represents an AUC of 0.5. The x-axis is the data set in which the model was initially trained on. The red lines represent the median AUC using that specific study as the training set.

Figure S6: Tissue Random Forest Prediction Success of Adenoma Using Genera

Across Studies. The red lines represent the median AUC using that specific study as the training set.

Figure S7: Most Common Genera Across Full Community Tissue Study Models. A) Common genera in the top 10 for adenoma Random Forest genus models. B) Common genera in the top 10 for unmatched carcinoma Random Forest genus models. B) Common genera in the top 10 for matched carcinoma Random Forest genus models. D) Common genera in the top 10 for adenoma Random Forest OTU models. E) Common genera in the top 10 for unmatched carcinoma Random Forest OTU models. F) Common genera in the top 10 for matched carcinoma Random Forest OTU models.