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

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 CRC by testing the hypothesis that the bacterial community has important associations to both early (adenoma) and late (carcinoma) stage disease. To test this hypothesis we examined both feces (n = 1737) and tissue (492 total samples from 350 individuals) across 14 different studies.

**Results.** Fecal samples had a significant decrease from control to adenoma to carcinoma for both Shannon diversity and evenness after correcting for study effect and variable region sequenced (P-value < 0.05). This reduction in evenness resulted in small increases in relative risk for adenoma (P-value = 0.032) and carcinoma (P-value = 0.00034) while the reduction in Shannon diversity only resulted in an increased relative risk for carcinoma (P-value = 0.0047). Previously associated colorectal cancer genera (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, or *Porphyromonas*) followed a similar pattern, with their presence significantly increasing the relative risk for carcinoma (P-value < 0.05) but not adenoma (P-value check tables 0.05) with the exception of *Porphyromonas* (P-value = 0.023). Using the whole community versus only CRC-associated genera to build a prediction model resulted in higher classification success, based on Area Under the Curve (AUC), for both adenoma and carcinoma using fecal and tissue samples. The most important OTUs for these models consistently belonged to genera such as *Ruminococcus*, *Bacteroides*, and *Roseburia* across studies. Overall, there were less associations between the microbiota and adenoma and one reason why this may be is that most studies were only adequately powered for large effect sizes.

**Conclusions.** This data provides support for the importance of the bacterial community to

both adenoma and carcinoma genesis. The evidence collected within this study on the role of the microbiota in CRC pathogenesis shows stronger associations between carcinoma then adenoma. One reason for this may be in part due to the low power to detect more subtle changes in the majority of studies that have been performed to date.

**Keywords**

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

## Background

Colorectal cancer (CRC) is a growing world-wide health problem [1] in which the microbiota has been purported to play an active role in disease pathogenesis [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 biomarkers or individual microbes but did not critically investigate how the community changes in CRC.

Although there has been an intense focus on microbiota based biomarker discovery for CRC, the candidate genera seem to be numerous. Some studies point towards mouth-associated genera such as *Fusobacterium*, *Peptostreptococcus*, *Parvimonas*, and *Porphyromonas* as key enriched genera [6,12–18]. Yet, even in some of these studies mouth-associated genera are far from the only microbes identified to be associated with CRC. These other genera include, but are not necessarily limited to, *Providencia*, *Mogibacterium*, *Enterococcus*, *Escherichia/Shigella*, *Klebsiella*, and *Streptococcus* [14–16]. In fact, *Escherichia/Shigella* and *Streptococcus* also have good *in vivo* evidence that individual species within the genus can be important in the pathogenesis of CRC [5,19,20]. Other studies have also identified *Akkermansia muciniphila* and *Bacteroides fragilis* as potential additional markers of CRC with *Bacteroides fragilis* having good mechanistic studies into how this may occur [15,21,22]. Still, other studies have identified genera like *Roseburia* as increased in CRC while others have reported it to be decreased or have found no difference in any genera at all [15,18,23,24].

Most of these studies have focused on carcinoma but the adenoma observations are not

any clearer at identifying candidate genera. Some groups have focused on more broad scale community metrics and have found that metrics such as richness are decreased in adenoma versus controls while others have identified *Lactococcus*, *Pseudomonas*, *Acidovorax*, *Cloacibacterium*, *Helicobacter*, *Lactobacillus*, *Bilophila*, *Desulfovibrio*, and *Mogibacterium* to be increased in adenoma [25–27]. A recent meta-analysis confirmed the correlations of certain mouth-associated genera and *Akkermansia muciniphila* with carcinoma [11]. However, it only analyzed a total of 509 stool samples and with some recent studies, that were not included, matching this sample size alone, it is hard to know how extrapolatable these findings truly 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*) that need to be investigated further, since a number of these genera have been found to be enriched in controls and not lesion (adenoma or carcinoma) [13,16,17].

Targeting the identification of biomarkers within stool seems logical since it offers an easy and cost-effective way to stratify risk and the current gold standard for diagnosis, a colonoscopy, can be time-consuming and is not risk free. Although stool represents an easy and less invasive way to assess risk, it is not clear how reflective this sample is to the community on the adenoma or carcinoma. Some studies have tried to assess this in health and disease but are limited by their sample size [18,28]. Sampling the tissue directly may provide clearer answers but is not without limitations. Due to bowel prep the communities left for sampling may not be reflective of the resident microbiota, but rather a collection of what is able to keep adhered to the mucosa. 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 end up being study dependent due to the randomness of contamination [29]. Due to these many differences that could arise between stool and tissue, one question our meta-analysis aims to answer

is whether there are consistent patterns that emerge across studies regardless of whether they used stool or tissue samples.

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 CRC. Importantly, this analysis and approach could provide valuable insights into the common genera that are both protective and detrimental in CRC and whether broad bacterial community measurements can account for these changes that were not previously provided by earlier meta-analysis studies [11].

Using both feces (n = 1737) and tissue (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 is an important risk factor or if select members can recapitulate the same information for both adenoma and carcinoma. To accomplish this we first assessed whether bacterial diversity changes throughout disease (control to adenoma to carcinoma) and if it results in an increased relative risk (RR) for adenoma or carcinoma. We then assessed what genera, if any, increase or decrease the RR of adenoma or carcinoma. Next, using Random Forest models, we analyzed whether the full community or only the top 5 increased and top 5 decreased RR genera combined resulted in better model classification, based on the area under the curve (AUC). Our results suggest that the community changes as disease severity worsens and that this community is important for disease classification. Finally, we also examined at what effect and sample size the studies that were used were powered for and the sample size needed to get to an 80% power. Our power analysis suggests that each individual study used was underpowered for detecting effect size differences of 10% or below between the case and control groups.

# Results

***Lower Bacterial Diversity is Associated with Increased RR of Carcinomas:*** Power transforming and Z-score normalizing OTU richness, evenness, and Shannon diversity for the entire data set allowed for the combination of data from all studies to assess differences in broad scale community metrics as disease severity worsens. 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 from control to adenoma to carcinoma (P-value = 0.025 and 0.043, respectively) [Figure 1A]. We next tested whether these detectable differences in community resulted in significant increases in RR of lesion (adenoma or carcinoma). For fecal samples, a decrease versus the overal median in evenness resulted in a significantly increased RR for carcinoma (RR = 1.36 (1.15 - 1.61), P-value = 0.00034) and adenoma (RR = 1.16 (1.01 - 1.34), P-value = 0.032) while a decrease versus the overal median in Shannon diversity only increased the RR for carcinoma (RR = 1.33 (1.09 - 1.62), P-value = 0.0047) [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 versus control and adenoma 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 variable region to test whether OTU richness, evenness, or Shannon diversity changed in a step-wise manner as disease severity increased. For tissue, there were no significant changes in OTU richness, evenness, and Shannon diversity from control to adenoma to carcinoma (P-value > 0.05) [Figure 1B & C]. When analyzing 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) tissue samples, for

both adenoma and carcinoma 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 also identified in unmatched tissue samples for both those with adenoma and carcinoma [Table S6 & S7]. For studies with matched samples no differences in bacterial community were observed when assessed with PERMANOVA [Table S6 & S7].

***Mouth-Associated Genera are Associated with an Increased RR of Lesion:*** Based on the small increase in RR using $\alpha$-diversity metrics, we assessed if being higher than the median for any specific genera resulted in an altered RR for adenoma or carcinoma in stool and tissue. To investigate this we analyzed all common genera across each study, in tissue or stool, and assessed whether a relative abundance higher than the median results in an increase or decrease in RR. For both stool and tissue none of the genera in the top 5 most significantly increased RR were the same between adenoma and carcinoma [Figure 3]. Mouth-associated genera were in the top 5 genera associated with an increased RR of adenoma (*Pyramidobacter* [Figure 3A] and *Rothia* [Figure 3C]) and carcinoma (*Fusobacterium*, *Parvimonas*, *Porphyromonas*, and *Peptostreptococcus* [Figure 3B] and *Fusobacterium* [Figure 3D]) for both stool and tissue. Conversely, genera commonly associated with a normal gastrointestinal tract were correlated with a decreased RR for both adenoma and carcinoma for both stool and tissue [Figure 3]. Overall, there was little direct overlap of the top 5 increased or decreased RR genera between stool and tissue.

For adenoma, there was almost no overlap between genera with a RR P-value of less than 0.05 and when they were similar the RR was in opposite directions (e.g. *Lactococcus*) [Table S8 & S9]. Many of the genera that had a P-value under 0.05 for either an increased or dereased RR are also highly prevalent in contamination, specifically, *Novosphingobium*, *Selemonas*, and *Achromobacter* [Figure 3 & Table S8-S9]. For carcinoma, certain

mouth-associated genera (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*) had a high RR for both tissue and stool [Table S10 & S11]. Although the most significant increased RR genera for tissue was *Camplyobacter* while in stool it was *Peptostreptococcus* [Table S10 & S11].

***Whole Community Models Add Important Context:*** Since specific genera increased RR for carcinoma over diversity metrics we assessed whether the bacterial community was better at classifying disease versus only a select group of genera based on RR and P-value significance and tested this using two approaches. The first approach used genus level data and tested whether there were any differences in AUC between all genera and selected genera during training on one study or testing on all other studies. The second approach used OTU level data and tested whether there was a generalized decrease in the 10-fold cross validation (CV) model AUC across studies when comparing all OTUs versus only OTUs that taxonomically classified to selected genera.

For the genera-based models the training set median AUC for model classification was similar for select genera and full genera models, for both tissue and stool studies [Figure S2-S3]. With respect to the test sets, comprised of genera data from other studies, both all and select genera models had a similar ability to detect adenomas or carcinomas, with the select genera models performing better in some instances [Figure S4-S6]. Conversely, the OTU-based models showed a slight decrease in median AUC between the full and select models, with one exception being carcinoma models for matched tissue [Figure 4 & 5].

In stool the most common genera in the full genera-based models belonged to genera such as *Ruminococcus*, *Bacteroides*, and *Roseburia* [Figure 6A & B]. Although all four CRC-associated genera were present in carcinoma, none were present in the majority of studies and *Fusobacterium* was the only genus present in adenoma [Figure 6A & B]. For the 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 stool, the tissue full genera-based Random Forest models showed no consistent representation of *Ruminococcaceae*, *Ruminococcus*, *Bacteroides*, and *Roseburia* in the top 10 across study [Figure S7]. The vast majority of the top 10 genera and OTUs occured in a study specific manner for these tissue based Random Forest models. For both the full genera and OTU Random Forest models of adenoma and carcinoma, there appears to be very little overlap in the top 10 most important variables between stool and tissue [Figure 6 & S7]. This discordance between stool and tissue also applies to the mouth-associated genera with one noticeable skew being that *Fusobacterium* and *Fusobacteriaceae* occur more often in the top 10 of Random Forest models using matched versus unmatched tissue samples [Figure S7B-C & S7E-F].

***CRC Studies are Underpowered for Detecting Small Effect Sizes:*** Based on our reported results we assessed whether the studies analyzed were realistically powered to identify small, medium, and large scale differences between cases and controls. When assessing the power of each study at different effect sizes the majority of studies, for both adenoma and carcinoma, acheived 80% power to detect a 30% or greater difference between groups [Figure 7A & B]. No study that was analyzed had the standard 80% power to detect an effect size difference that was 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 each arm [Figure 7C].

# Discussion

Our study identifies clear differences in diversity, both at the community level and for individual genera, that are present in patients with and without CRC [Figure 1-3]. Although there was a step-wise decrease in diversity from control to adenoma to carcinoma, this did not translate into large effect sizes for the RR of lesion. Even though mouth-associated genera increased the RR of carcinoma, they did not consistently increase the RR of adenoma. These mouth-associated genera are clearly important to carcinoma classification but our observations suggest that accounting for the community in which these microbes exist can increase the ability of models to make predictions.

The data presented herein supports the importance of select genera for carcinoma but not necessarily adenoma formation. Our observations show that when only using the select genera to create models for carcinoma the AUC are similar to using all genera or OTUs [Figure 4 & 5]. This suggests that an interplay between a select number of potentially protective and exacerbating microbes, within the GI community, are crucial for carcinoma formation. Importantly, it suggests that there may be key members of the 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 need to be used to identify members of the community that are associated with adenoma. 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 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 potential hypothesis from these observations is that at early stages of the disease, how the host interacts with these subtle changes is what ultimately 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 study [Figure 6]. Changes in *Bacteroides*, *Ruminococcaceae*, *Ruminococcus*, and *Roseburia* were consistently found to be disciminative across the different studies for both adenoma and carcinoma [Figure 6]. This data would suggest that whether the non-resident bacterium is *Fusobacteria* or *Peptostreptococcus* is not 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.

The tissue studies did not provide a clearer understanding of how the microbiota may be associated with lesion. Generally, the full OTU-based models of unmatched and matched [Figure S7E & F] tissue samples were concordant with stool showing that resident microbes were the most prevalent in the top 10 across study. Unlike in stool, *Fusobacterium* was the only mouth-associated bacteria consistently present in the top 10 of the full CRC models [Figure S7B-C & E-F]. The majority of the tissue results seem to be study specific with many 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 because it is commonly a marker of contamination. The low bacterial biomass of tissue samples coupled with potential contamination could ultimately explain why the these results seem to be more sporadic than the stool results.

One important caveat to this study is that even though specific genera associated with species such as *Bacteroides fragilis* and *Streptococcus gallolyticus* subsp. *gallolyticus* were not identified, it does not necessarily mean these specific species are not important in human CRC [20,22]. Since we are limited in our aggregation of the data, across study, to the genus level, it is not possible to clearly delineate what specific 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, for stool, the full RF models consistently identified the genus *Bacteroides* as well as OTUs that classified as *Bacteroides* to be important model component across studies. Suggesting that even though *Bacteroides* may not increase the RR of lesion and may not vary in relative abundance, like *Fusobacterium*, it is still important in CRC. Additionaly, *Streptococcus gallolyticus* subsp. *gallolyticus* is a mouth-associated micobe and our study clearly shows that regardless of whether the sample is tissue or stool, mouth-associated genera are commonly associated with an increased RR of both adenoma and carcinoma.

The associations between the microbiota and adenoma 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. A small effect size may well be the scope in which differences consistently occur in adenoma based on the results observed within our meta-analysis. Future studies investigating adenoma and the microbiota need to take power into consideration if we are to reproducibly study whether the microbiota contributes to poly formation. In contrast to adenoma, our observations suggest that most studies analyzed have sufficient power to detect many changes in carcinoma because of large effect size differences between cases and controls [Figure 7].

## Conclusion

By aggregating together a large collection of studies from both feces and tissue, we are able to provide evidence in support of the importance of the bacterial community in CRC. Further, the data presented here suggests that mouth-associated microbes can gain a foothold within the colon and are are commonly associated with the greatest RR of carcinoma. Conversely, no conclusive signal with these mouth-associated micobes could be detected for adenoma. Our observations also highlight the importance of power and sample number considerations when undertaking investigations into the microbiota and adenoma due to the subtle changes in the community. Overall, the microbiota associations with carcinoma are much stronger than the associations with adenoma.

## 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 were included and only data sets that had 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 and were excluded. 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 not used which left 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 that were 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 was 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, used in mothur, for either 454 or Illumina based sequencing. If it was not possible to use the defaults, the stated quality cut-offs, from the study itself, were used instead. 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 the $\alpha$-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 $\alpha$-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 hypervariable 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. $\beta$-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:*** $\alpha$-diversity was first assessed for differences between controls, adenoma, and carcinoma using both linear mixed-effect models and RR. $\beta$-diversity was then assessed for each individual study for differences between control-adenoma and control-carcinoma. Next, all common genera were assessed for differences in RR for

adenoma and carcinoma and ranked based on P-value. We then built Random Forest models based on all or a selected model, based on the top 5 increased and top 5 decreased RR based on P-value, and these models were trained on one study then tested on the remaining studies, for every study. A similar approach was then applied at the OTU level with the exception that a 10-fold CV over 100 different models, 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 variable in the top 5 increased and top 5 decreased RR 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 normal versus adenoma were made the carcinoma samples were excluded from each respective study. Similarly, for comparisons in which normal versus carcinoma were made the adenoma samples were excluded from each respective study. The data was split between feces and tissue samples. Within the tissue groups the data was further divided between samples from the same individual (matched) and those from different individuals (unmatched). Where applicable for each study, predictions for adenoma and carcinoma were then tested for feces, matched tissue, and unmatched tissue.

***Obtaining Genera Relative Abundance and Selected Models:*** For the genera analysis of the RR, OTUs were added together based on the genus or lowest available taxonomic classification level and the total average counts, for 100 different subsamplings, were collected. The OTU based Random Forest Models using selected OTUS utilized a similar approach except that the OTUs were not aggregated together by taxonomic identity but kept as separate OTUs. OTU Random Forest models using the full community included all OTUs while those for the selected model included only those OTUs that had a taxonomic classification to a variable in the top 5 increased of top 5 decreased RR based on P-value.

***Matched versus Unmatched Tissue Samples:*** In general, tissue samples that had control and lesion samples, that did not belong to the same individual, 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 due to one of the corresponding matched samples not making 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 were obtained in two different ways depending on whether the model used genera or OTU data. For the 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 while, for 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. The two studies that had adenoma tissue were equally divided between matched and unmatched groups and were grouped together for the counting of the top 10 genera and OTUs.

***Reproducible Methods:*** The code and analysis can be found here at https://github.com/SchlossLab/Sze_CRCMetaAnalysis_Microbiome_2017. Unless otherwise mentioned, the accession number for the raw sequences for the studies used in this analysis can be found

391 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, N.Y.). 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.

2014;10:766.

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

19. 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 [Internet]. Springer Nature; 2014;5:4724. Available from: https://doi.org/10.1038/ncomms5724

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

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

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

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

24. 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.

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. Appl.Environ.Microbiol. [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

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

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

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

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

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

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

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

552 46. Liaw A, Wiener M. Classification and regression by randomForest. R News [Internet].
553 2002;2:18–22. Available from: http://CRAN.R-project.org/doc/Rnews/

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

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

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

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

50. Auguie B. GridExtra: Miscellaneous functions for “grid” graphics [Internet]. 2017. 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: $\alpha$-Diversity Differences between Control, Adenoma, and Carcinoma Across Sampling Site.** A) $\alpha$-diversity metric differences by group in stool samples. B) $\alpha$-diversity metric differences by group in unmatched tissue samples. C) $\alpha$-diversity metric differences by group in matched tissue samples. The dashed line represents a Z-score of 0 or no difference from the median.

**Figure 2: Relative Risk for Adenoma or Carcinoma based on $\alpha$-Diversity Metrics in Stool.** A) $\alpha$-metric relative risk for adenoma. B) $\alpha$-metric relative risk for carcinoma. Colors represent the different variable regions used within the respective study.

**Figure 3: Top 5 Genera that Decrease and Increase Relative Risk for Lesion.** A) Adenoma relative risk in stool. B) Carcinoma relative risk in stool. C) Adenoma relative risk in tissue. D) Carcinoma relative risk in tissue. For all panels the relative risk was also compared to whether one, two, three, or four of the CRC-associated genera were present.

**Figure 4: Stool OTU Random Forest Model Across Studies.** A) Adenoma random forest model between the full community and select genera OTUs only. B) Carcinoma random forest model between the full community and select genera 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 5: Tissue OTU Random Forest Model Across Studies.** A) Adenoma random forest model between the full community and select genera OTUs only. B) Carcinoma random forest model between the full community and select genera 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: Relative Risk for Adenoma or Carcinoma based on $\alpha$-Diversity Metrics in Tissue.** A) $\alpha$-metric relative risk for adenoma. B) $\alpha$-metric relative risk 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.