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 both individual and collections of bacteria are associated with the progression of colorectal cancer. As the number of studies investigating these associations increases and the number of subjects in each study increases, a meta-analysis to identify the associations that are the most predictive of disease progression is warranted. For our meta-analysis, we analyzed previously published 16S rRNA gene sequencing data collected from feces (1737 individuals from 8 studies) and colon tissue (492 total samples from 350 individuals from 7 studies).

Results. We quantified the odds ratios for individual bacterial genera that were associated with an individual having tumors relative to a normal colon. Among the stool samples, there were no genera that had a significant odds ratio associated with adenoma and there were 8 genera with significant odds ratios associated with carcinoma. Similarly, among the tissue samples, there were no genera that had a significant odds ratio associated with adenoma and there were 3 genera with significant odds ratios associated with carcinoma. Among the significant odds ratios, the association between individual taxa and tumor diagnosis was equal or below 7.11. Because individual taxa had limited association with tumor diagnosis, we trained Random Forest classification models using the genera with the five highest and lowest odds ratios, using the entire collection of genera found in each study, and using operational taxonomic units defined based on a 97% similarity threshold. All training approaches yielded similar classification success as measured using the Area Under the Curve. The ability to correctly classify individuals with adenomas was poor and the ability to classify individuals with carcinomas was considerably better using sequences from stool or tissue.

Conclusions. This meta-analysis confirms previous results indicating that individuals with adenomas cannot be readily classified based on their bacterial community, but that those

with carcinomas can. Regardless of the dataset, we found a subset of the fecal community that was associated with carcinomas was as predictive as the full community.

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 hypothesized to have a role in disease progression [1,2]. Numerous studies using murine models of CRC have shown the importance of both individual microbes [3–7] and the overall community [8–10] in tumorigenesis. Numerous case-control studies have characterized the microbiota of individuals with colonic adenomas and carcinomas in an attempt to identify biomarkers of disease progression [6,11–17]. Because current CRC screening recommendations are poorly adhered to due to socioeconomic status, test invasiveness, and frequency of tests, development and validation of microbiome-associated biomarkers for CRC progression could further attempts to develop non-invasive diagnostics [18].

Recently, there has been an intense focus on identifying microbiota-based biomarker yielding a seemingly endless number of candidate taxa. Some studies point towards mouth-associated genera such as Fusobacterium, Peptostreptococcus, Parvimonas, and Porphyromonas that are enriched in people with carcinomas [6,11–17]. Other studies have identified members of *Akkermansia*, *Bacteroides*, *Enterococcus*, *Escherichia*, *Klebsiella*, *Mogibacterium*, *Streptococcus*, and *Providencia* are also associated with carciomas [13–15]. Additionally, *Roseburia* has been found in some studies to be more abundant in people with tumors but in other studies it has been found to be either less abundant or no different than what is found in subjects with normal colons [14,17,19,20]. There are strong results from tissue culture and murine models that Fusobacterium nucleatum, pks-positive strains of Escherichia coli, Streptococcus gallolyticus, and an entertoxin-producing strain of Bacteroides fragilis are important in the pathogenesis of CRC [5,14,21–24]. These results point to a causative role for the microbiota in CRC pathogenesis as well as their potential as diagnostic biomarkers.

Most studies have focused on identifying biomarkers in patients with carcinomas but there is a greater clinical need to identify biomarkers associated with adenomas. Studies focusing on broad scale community metrics have found that measures such as the total number of Operational Taxonomic Units (OTUs) are decreased in those with adenomas versus controls [25]. Other studies have identified *Acidovorax*, *Bilophila*, *Cloacibacterium*, *Desulfovibrio*, *Helicobacter*, *Lactobacillus*, *Lactococcus*, *Mogibacterium*, and *Pseudomonas* to be enriched in those with adenomas [25–27]. There are few genera that are enriched in patients with adenoma or carcinoma tumors.

Confirming some of these previous findings, a recent meta-analysis found that 16S rRNA gene sequences from members of the *Akkermansia*, *Fusobacterium*, and *Parvimonas* were fecal biomarkers for the presence of carcinomas [28]. Contrary to previous studies they found sequences similar to members of *Lactobacillus* and *Ruminococcus* to be enriched in patients with adenoma or carcinoma relative to those with normal colons [12,15,16]. In addition, they found 16S rRNA gene sequences from members of *Haemophilus*, *Methanosphaera*, *Prevotella*, *Succinovibrio* were enriched in patients with adenoma and *Pantoea* were enriched in patients with carcinomas. Although this meta-analysis was helpful for distilling a large number of possible biomarkers, the aggregate number of samples included in the analysis (n = 509) was smaller than several larger case-control studies that have since been published [12,27]

Here we provide an updated meta-analysis using 16S rRNA gene sequence data from both feces (n = 1737) and colon tissue (492 samples from 350 individuals) from 14 studies [11–17,19,20,23,25–27,29] [Table 1 & 2]. We expand both the breadth and scope of the previous meta-analysis to investigate whether biomarkers describing the bacterial community or specific members of the community can more accurately classify patients as having adenoma or carcinoma. Our results suggest that the bacterial community changes as disease severity worsens and that that a subset of the microbial community can be

used to diagnose the presence of carcinoma.

Results

Lower Bacterial Diversity is Associated with Increased Odds Ratio (OR) of Tumors: We first assessed whether variation in broad community metrics like total number of operational taxonomic units (OTUs) (i.e. richness), the evenness of their abundance, and the overall diversity was associated with disease stage after controlling for study and variable region differences. In stool, there was a significant decrease in both evenness and diversity as disease severity progressed from normal to adenoma to carcinoma (P-value = 0.025 and 0.043, respectively) [Figure 1]; there was not a significant difference for richness (P-value = 0.21). We next tested whether the decrease in these community metrics translated into significant ORs for having an adenoma or carcinoma. For fecal samples, the ORs for richness were not significantly greater than 1.0 for adenoma or carcinoma (P-value = 0.40) [Figure 2A]. The ORs for evenness were significantly higher than 1.0 for adenoma (OR = 1.3 (1.02 - 1.65), P-value = 0.035) and carcinoma (OR = 1.66 (1.2 - 2.3), P-value = 0.0021) [Figure 2B]. The ORs for diversity were only significantly greater than 1.0 for carcinoma (OR = 1.61 (1.14 - 2.28), P-value = 0.0069), but not for adenoma (P-value = 0.11) [Figure 2C]. Although these OR are significantly greater than 1.0, it is doubtful that these are clinically meaningful values.

Similar to our analysis of sequences obtained from stool samples, we repeated the analysis using sequences obtained from colon tissue. There were no significant changes in richness, evenness, or diversity as disease severity progressed from control to adenoma to carcinoma (P-value > 0.05). We next analyzed the OR, for matched (i.e. where unaffected tissue and tumors were obtained from the same individual) and unmatched (i.e. where unaffected tissue and tumor tissue were not obtained from the same individual) tissue samples. The ORs for adenoma and carcinoma by any measure were not significantly different from 1.0 (P-value > 0.05) [Figure S1 & Table S1]. This is likely due to the combination of a small effect size, as suggested from the results using stool, and the

relatively small number of studies and size of studies used in the analysis.

Disease Progression is Associated with Community-Wide Changes in Composition and Abundance: Based on the differences in evenness and diversity, we next asked whether there were community-wide differences in the structure of the communities associated with different disease stages. We identified significant bacterial community differences in the stool of patients with adenomas relative to those with normal colons in 1 of 4 studies and in patients with carcinomas relative to those with normal colons in 6 of 7 studies (PERMANOVA; P-value < 0.05) [Table S2]. Similar to the analyses using stool samples, there were significant differences in bacterial community structure between subjects with normal colons and those with adenoma (1 of 2 studies) and carcinoma (1 of 3 studies) [Table S2]. For studies that used matched samples no differences in bacterial community structures were observed [Table S2]. Combined, these results indicate that there consistent and significant community-wide changes in the fecal community structure of subjects with carcinomas. However, the signal observed in subjects with adenomas or when using tissue samples was not as consistent. This is likely due to a smaller effect size or the relatively small sample sizes among the studies that characterized the tissue microbiota.

Individual Taxa are Associated with Significant ORs for Carcinomas: Next we identified those taxa were associated with ORs that were significantly associated with having a normal colon or the presence of adenomas or carcinomas. No taxa had a significant OR for the presence of adenomas when we used data collected from stool or tissue samples (Table S3 & S4). In contrast, 8 taxa had significant ORs for the presence of carcinomas using data from stool samples. Of these, 4 are commonly associated with the oral cavity: Fusobacterium (OR = 2.74 (1.95 - 3.85)), Parvimonas (OR = 3.07 (2.11 - 4.46)), Porphyromonas (OR = 3.2 (2.26 - 4.54)), and Peptostreptococcus (OR = 7.11 (3.84 - 13.17)) [Table S3]. The other 4 were Clostridium XI (OR = 0.65 (0.49 - 0.86)),

Enterobacteriaceae (OR = 1.79 (1.33 - 2.41)), Escherichia (OR = 2.15 (1.57 - 2.95)), and Ruminococcus (OR = 0.63 (0.48 - 0.83)). Among the data collected from tissue samples, only unmatched carcinoma samples had taxa with a significant OR. Those included Dorea (OR = 0.35 (0.22 - 0.55)), Blautia (OR = 0.47 (0.3 - 0.73)), and Weissella (OR = 5.15 (2.02 - 13.14)). Mouth-associated genera were not significantly associated with an increased OR for carcinoma in tissue samples [Table S4]. For example, Fusobacterium had an OR of 3.98 (1.19 - 13.24; however, due to the small number of studies and considerable variation in the data, the Benjimani-Hochberg-corrected P-value was 0.93 [Table S4]. It is interesting to note that Ruminococcus and members of Clostridium group XI in stool and Dorea and Blautia in tissue had ORs that were significantly less than 1.0, which suggests that these populations are protective against the development of carcinomas. Overall, there was no overlap in the taxa with significant OR between stool and tissue samples.

Individual taxa with a significant OR do a poor job of differentiating subjects with normal colons and those with carcinoma: We next asked whether those taxa that had a significant OR associated with having a normal colon or carcinomas could be used individually, to classify subjects as having a normal colon or carcinomas. Whereas the OR was defined based on whether the relative abundance for a taxon in a subject was above or below the median relative abundance for that taxon across all subjects in a study, we generated receiver operator characteristic (ROC) curves for each taxon in each study and calculated the area under the curve (AUC). This allowed us to use a more fluid relative abundance threshold for defining disease status. Using data from stool samples, the 8 taxa did no better at classifying the subjects than one would expect by chance (i.e. AUC=0.50) [Figure 3A]. The taxa that performed the best included Clostridium XI, Ruminococcus, and Escherichia and even these had median AUC values less than 0.588. Likewise, in unmatched tissue samples the 8 taxa with significant ORs taxa were marginally better than one would expect by chance [Figure 3B]. The relative abundance of Dorea was the best predictor of carcinomas and its median AUC was only 0.62. These results suggest that

although these taxa are associated with a decreased or increased OR for the presences of carcinomas, individually, they do a poor job of classifying a subject's disease status.

Combined taxa model classifies subjects better than using individual taxa: Instead of attempting to classify subjects based on individual taxa, next we generated Random Forest models that combined the individual taxa and evaluated the ability to classify as subject's disease status. For data from stool samples, the combined model had an AUC of 0.75, which was significantly higher than any of the AUC values for the individual taxa (P-value < 0.033). For the full taxa models using stool, *Bacteroides* and *Lachnospiraceae* were the most common taxa in the top 10% mean decrease in accuracy (MDA) across studies. Similarly, using data from the unmatched tissue samples, the combined model had an AUC of 0.77, which was significantly higher than the AUC values for Blautia and Weissella (P-value < 0.037). For the full taxa models using unmatched tissue, *Lachnospiraceae*, *Bacteroidaceae*, and *Ruminococcaceae* were the most common taxa in the top 10% mean decrease in accuracy across studies. Clearly, pooling the information from the taxa with significant ORs results in a model that outperforms classifications made using individual taxa.

Performance of models based on taxa relative abundance in full community are not significantly better than those based on taxa with significant ORs: Next, we asked whether a Random Forest classification model built using all of the taxa found in the communities would outperform the models generated using those taxa with a significant OR. Similar to our inability to identify taxa associated with a significant OR for the presence of adenomas, the median AUCs to classify subjects as having normal colons or having adenomas using data from stool or tissue samples were marginally better than 0.5 for any study[Figure 4A & S3A]. In contrast, the models for classifying subjects as having normal colons or having carcinomas using data from stool or tissue samples yielded AUC values meaningfully higher than 0.5 [Figure 4B & S3B-C]. When we compared the models based

on all of the taxa in a community to models based on the taxa with significant ORs, the results were mixed. Using the data from stool samples we found that although the AUC for 6 of 7 studies increased (mean decrease = 9.53%), the more expansive models performed worse for 1 of the studies (decrease = 0.38%). The overall improvement in performance was statistically significant (one-tailed paired T-test; P-value = 0.005). Of the 8 taxa with significant ORs, all 8 were among the top 10% most important taxa as measured by mean decrease in accuracy, in at least one study . Similarly, using the data from unmatched tissue samples we found that the AUC for 4 out of 4 studies decreased between full versus select OR models (mean decrease = 19.11%, one-tailed paired T-test; P-value = 0.03). Of the 3 taxa with significant ORs, all 3 were among the top 10% most important taxa as measured by mean decrease in accuracy, in at least one study. These results were surprising because it demonstrated that the ability to classify subjects could be done based on a limited characterization of the communities.

Performance of models based on OTU relative abundance in full community are not significantly better than those based on taxa with significant ORs: The previous models were based on relative abundance data where sequences were assigned to coarse taxonomic assignments (i.e. typically genus or family level). To determine whether model performance improved with a more fine scale classification, we assigned sequences to operational taxonomic units (OTUs) where the similarity among sequences within an OTU was more than 97%. We again found that classification models built using all of the sequence data for a community did a poor job of differentiating between subjects with normal colons and those with adenomas (median AUC: 0.53 [0.37- 0.56]), but did a good job of differentiating between subjects with normal colons and those with carcinomas (median AUC: 0.71 [0.5- 0.9]). The OTU-based models performed similarly to those constructed using the taxa with significant ORs (one-tailed paired T-test; P-value = 0.966) and those using all taxa (one-tailed paired T-test; P-value = 0.146). Among the OTUs that had the highest mean decrease in accuracy for the OTU-based models, we found that

OTUs that affiliated with all of the 8 taxa that had a significant OR were within the top 10% for at least one study. Again, this result was surprising as it indicated that a finer scale classification of the sequence data and thus a larger number of features to select from, did not yield improved classification of the subjects.

Generalizability of taxon-based models trained on one dataset to the other datasets: Considering the good performance of the Random Forest models using taxa with a significant OR and using all of the taxa, we next asked how well the models would perform when given data from a different subject cohort. For instance, if a model was trained using data from the Ahn study, we wanted to know how well it would perform using the data from the Baxter study. We found the models trained using the taxa with a significant OR all had a higher median AUC than the models trained using all of the taxa when tested on the other datasets [Figure 5]. As might be expected, the difference between the performance of the modelling approaches appeared to vary with the size of the training cohort. These data suggest that given a sufficient number of subjects with normal colons and carcinomas, Random Forest models trained using a small number of taxa can accurately classify individuals from a different cohort.

Discussion

Select subset community you can do as well as OTUs or genera when using all of the information and that these models perform well when trained on one cohort and tested on another.

Why didn't the whole community work better? Communities are patchy and higher level taxonomic information pools some of that patchiness There isn't one single bug associated with disease – why ? possibly because multiple bugs can carry out the same function (e.g. inflammation). Tell some examples (Esch Fuso, pepto) Also, interesting that the loss of bacteria can be associated with disease. These appear to be commonly thought as beneficial bacteria that ferment to produce SCFAs

Adenomas suck. Which is kind of at odds with Baxter, but Baxter looked at lesions and also included FIT Also, variation in how adenomas are defined – small polyps vs larger neopplasias (e.g. baxter), but we couldn't do this for the current study because we didn't have that level of detail and most studies are small This isn't such a novel result (Nich Chia) Stool may not be the best place to look for early stage adenomas

Tissue vs. Stool Not much overlap between stool and tissue biomarkers Stool for diagnostic – does as well as tissue for classifying carcinomas Carcinoma result surprising (e.g. lack of Fuso) – largely because small studies?

Missing microbes ETBfrag, Strepto, etc. B – is a lot of things that we might lose – good and bad; mucosal, right sided tumors? We might not see with stool

Meta analysis are important (eg. obesity, crc), we need more. We learned that these are hard because... Inaccessible datasets Missing/vague metadata Varying sequence quality Small datasets

Looking forward Potential for using microbiome as a diagnostic tool We're excited that we were able to validate a set of biomarkes across multiple studies Found a phenotype we can relate to the microbiome, it's replicable and it's strong

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 [17,30]. 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 [31].

Our study identifies clear but small differences in diversity at the community level and larger differences for individual genera, present in patients with tumors versus controls [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 size increases in OR for either adenoma or carcinoma tumors. Even though mouth-associated genera increased individaul's OR of having a carcinoma for certain sample types, they did not consistently increase the OR of having an adenoma. By using these taxa that had significant ORs after multiple comparison correction we found that we could classify indviduals with either adenoma or carcinoma as well as models that use either all genera or all OTUs. Finally, many studies were individually under powered to be able to reject the null hypothesis and

this could one reason only the comparison between control and carcinoma individuals for stool samples had relible detectable differences.

The data presented herein support the importance of specific taxa for carcinoma, but not necessarily adenoma, tumor formation. The results that we have presented show that the significant OR taxa model and both the full genera and OTU models, for indviduals with carcinoma, had similar AUCs [Figure 2 & 3]. This suggests that an interplay between a select number of potentially protective and exacerbating microbes within the GI community could be crucial for carcinoma formation. Importantly, it suggests that there may be key members of the GI community that should be studied further to potentially help reduce the risk of carcinoma tumor formation. Conversely, using the present data, it is clear that new approaches may be needed to identify members of the community associated with adenoma tumors. Regardless of sample type and whether a full genera- or OTU-based 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. This change in diversity, at this early stage of disease, could be focal to the adenoma itself. How the host interacts with these subtle changes at early stages of the disease could be what leads to a thoroughly dysfunctional community that is supportive of tumorgenesis.

For the full genera- and OTU-based models within stool, common GI microbes were most consistently present in the top 10 genera or OTUs across studies [Figure 4]. 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 individuals with adenoma and carcinoma [Figure 4]. 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 or general inflammatory 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 S4E & 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 S4B-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 [22,24]. 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 OR of individuals having an adenoma or carcinoma 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 OR for individuals to have a carcinoma tumor.

The associations between the microbiota and individuals with adenoma tumors 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 those with adenomas. Future studies investigating adenoma tumors and the microbiota need to take power into consideration to reproducibly study whether the microbiota contributes to adenoma 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 5].

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. 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. Conversely, no conclusive signal with these mouth-associated microbes could be detected for individuals with an adenoma. Our observations also highlight the importance of power and sample number considerations when investigating the microbiota and adenoma tumors due to possible subtle changes in the community. Overall, the associations between the microbiota and individuals with carcinomas were much stronger than with those with adenomas.

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.* [32,33] 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.* [11], Baxter, *et al.* [12], Brim, *et al.* [29], Burns, *et al.* [15], Chen, *et al.* [13], Dejea, *et al.* [20], Flemer, *et al.* [17], Geng, *et al.* [19], Hale, *et al.* [27], Kostic, *et al.* [34], Lu, *et al.* [26], Sanapareddy, *et al.* [25], Wang, *et al.* [14], Weir, *et al.* [23], and Zeller, *et al.* [16]. The Zackular [35] study was not included because the 90 individuals analyzed within the study are contained within the larger Baxter study [12]. 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, [11]) and directly from the authors (n = 4, [17,23,25,27]). Each study was processed using the mothur (v1.39.3) software program [36] 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 [37] before *de novo* OTU clustering at 97% similarity was completed using the OptiClust algorithm [38].

Study Analysis Overview: OTU richness, evenness, and Shannon diversity were first assessed for differences between controls, adenoma tumors, and carcinoma turmors using both linear mixed-effect models and ORs. For each individual study the Bray-Curtis index was used to assess differences between control-adenoma and control-carcinoma individuals. Next, all common genera were assessed for differences in ORs for individuals having an adenoma or carcinoma and corrected for multiple comparisons using the Benjamini-Hochberg method [39]. We then built Random Forest models based on all genera, all OTUs, or significant OR taxa (only using taxa still significant after multiple comparison correction). For both the full genera and significant OR taxa, models were trained on one study then tested on the remaining studies using genera-based relative abundances. The OTU-based 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. This process was repeated for every study in the meta-analysis. Comparisons of the initial trained model AUCs for the full genera and significant OR taxa were made to the mean AUC generated from the 100 different 10-fold CV runs of the respective OTU-based model. 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 fecal 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 Significant OR Taxa Models: For the genera analysis of the ORs, OTUs were added together based on the genus or lowest available taxonomic classification level and the total average counts, for 100 different subsamplings was obtained. The significant OR taxa models for the Random Forest models utilized all taxa that had significant ORs after multiple comparison correction. This meant only models for the carcinoma stool (8 variables) and carcinoma unmatched (3 variables) samples were possible to be created and analyzed.

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 matched and unmatched tissue samples, were analyzed together for the linear mixed-effect models with samples from the same individual being corrected for. All other analysis, where it is not specified explicitly, 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 specific taxa 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 were then identified by using the lowest classification for each of the specific OTUs obtained from these counts and the number of times this classification occurred across the top 10 of each study was recorded. Finally, the two studies that had adenoma tumor 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 for both Random Forest approaches.

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

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

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

515 9. Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, et al. The gut
516 microbiome modulates colon tumorigenesis. mBio. 2013;4:e00692–00613.

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

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

521 12. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model improves the
522 sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine.
523 2016;8:37.

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

526 14. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut
527 microbiota between colorectal cancer patients and healthy volunteers. The ISME journal.
528 2012;6:320–9.

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

531 16. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of
532 fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology.
533 2014;10:766.

534 17. Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, et al.
535 Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut.

2017;66:633–43.

18. García AZG. Factors influencing colorectal cancer screening participation. Gastroenterology Research and Practice [Internet]. Hindawi Limited; 2012;2012:1–8. Available from: https://doi.org/10.1155/2012/483417

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

20. 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.

21. 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

22. 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

23. 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.

24. 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);

559 2014;60:208–15. Available from: https://doi.org/10.1093/cid/ciu787

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

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

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

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

572 29. Brim H, Yooseph S, Zoetendal EG, Lee E, Torralbo M, Laiyemo AO, et al. Microbiome
573 analysis of stool samples from African Americans with colon polyps. PloS One.
574 2013;8:e81352.

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

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

s12915-014-0087-z

32. 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

33. 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

34. 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.

35. 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.

36. 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

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

38. Westcott SL, Schloss PD. OptiClust, an Improved Method for Assigning

603 Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere. 2017;2.

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

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

610 41. Mangiafico S. Rcompanion: Functions to support extension education program
611 evaluation [Internet]. 2017. Available from: https://CRAN.R-project.org/package=
612 rcompanion

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

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

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

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

624 46. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Vegan:
625 Community ecology package [Internet]. 2017. Available from: https://CRAN.R-project.org/

626 package=vegan

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

630 48. Liaw A, Wiener M. Classification and regression by randomForest. R News [Internet].
631 2002;2:18–22. Available from: http://CRAN.R-project.org/doc/Rnews/

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

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

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

Study	Data Stored	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	Region	Control (n)	Adenoma (n)	Carcinoma (n)
Burns	SRA	V5-6	18	0	16
Chen	SRA	V1-3	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: Significant Bacterial Community Metrics for Adenoma or Carcinoma in Stool. A) Adenoma evenness. B) Carcinoma evenness. C) Carcinoma Shannon diversity. Blue represents controls and red represents either adenoma (panel A) or carcinoma (panel B and C). The black lines represent the median value for each repsective group.

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: The AUC of Indivdiual Significant OR Taxa to classify Carcinoma. A) Stool samples. B) Unmatched tissue samples. The larger circle represents the median AUC of all studies and the smaller circles represent the individual AUC for a particular study. The dotted line denotes an AUC of 0.5.

Figure 4: Stool Random Forest Model Train AUCs. A) Adenoma random forest model AUCs between all genera, all OTU, and select model based on significant OR taxa. B) Carcinoma random forest model AUCs between all genera, all OTU, and select model based on significant OR taxa. The black line represents the median AUC for the respective group. If no values are present in the singificant OR taxa group then there were no significant taxa identified and no model was tested.

Figure 5: Stool Random Forest Genus-Based Model Test AUCs. A) Test AUCs of adenoma models using all genera across study. B) Test AUCs of carcinoma models using all genera or significant OR taxa only. The black line represents the AUC at 0.5. The red lines represent the median AUC of all test AUCs for a specific study.

Figure 6: Most Common Taxa Across Carcinoma Full Community Stool Study Models. A) Common taxa in the top 10 percent for carcinoma Random Forest all taxa-based models. B) Common taxa in the top 10 percent for carcinoma Random Forest

663 all OTU-based models.

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: Tissue Random Forest Model Train AUCs. A) Adenoma random forest model AUCs between all genera, all OTU, and select model based on significant OR taxa in unmatched and matched tissue. B) Carcinoma random forest model AUCs between all genera, all OTU, and select model based on significant OR taxa in unmatched and matched tissue. The black line represents the median AUC for the respective group. If no values are present in the singificant OR taxa group then there were no significant taxa identified and no model was tested.

Figure S3: Tissue Random Forest Genus-Based Model Test AUCs. A) Test AUCs of adenoma models using all genera across study. B) Test AUCs of carcinoma models using all genera for matched tissue studies. C) Test AUCs of carcinoma models using all genera or significant OR taxa only for unmatched tissue studies The black line represents the AUC at 0.5. The red lines represent the median AUC of all test AUCs for a specific study.

Figure S4: Most Common Genera Across Full Community Tissue Study Models. A) Common genera in the top 10 percent for matched carcinoma Random Forest all genera-based models. B) Common genera in the top 10 percent for unmatched carcinoma Random Forest all genera-based models. C) Common genera in the top 10 percent for matched carcinoma Random Forest all OTU-based models. D) Common genera in the top 10 percent for unmatched carcinoma Random Forest all OTU-based models.