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

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. We analyzed previously published 16S rRNA gene sequencing data collected from feces and colon tissue.

We quantified the odds ratios (ORs) for individual bacterial taxa that were associated with an individual having tumors relative to a normal colon. Among the fecal samples, there were no taxa that had a significant ORs associated with adenoma and there were 8 taxa with significant ORs associated with carcinoma. Similarly, among the tissue samples, there were no taxa that had a significant OR associated with adenoma and there were 3 taxa with significant ORs associated with carcinoma. Among the significant ORs, 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 only the taxa that had significant ORs, using the entire collection of taxa 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 fecal or tissue.

Importance

Colorectal cancer is a significant and growing health problem in which animal models and epidemiological data suggest that the colonic microbiota have a role in tumorigenesis. These observations indicate that the colonic microbiota is a reservoir of biomarkers that may improve our ability to detect colonic tumors using non-invasive approaches. This meta-analysis identifies and validates a set of 8 bacterial taxa that can be used within a Random Forest modeling framework to differentiate individuals as having normal colons or carcinomas. When models trained using one dataset were tested on other datasets, the models performed well. These results lend support to the use of fecal biomarkers for the detection of tumors. Furthermore, these biomarkers are plausible candidates for further mechanistic studies into the role of the gut microbiota in tumorigenesis.

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 person's socioeconomic status, test invasiveness, and frequency of tests, development and validation of microbiota-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 biomarkers 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* (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 less abundant than what is found in subjects with normal colons (14, 17, 19, 20). There is support from mechanistic studies using 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 tumorigenesis (5, 14, 21–24). These results point to a causative role for the microbiota in tumorigenesis as well as their potential as diagnostic biomarkers.

Most studies have focused on identifying biomarkers in patients with carcinomas but

there is a clinical need to identify biomarkers associated with adenomas to facilitate early detection of the tumors. Studies focusing on broad scale community metrics have found that measures such as the total number of taxa (i.e. richness) are lower 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). The ability to classify individuals as having normal colons or adenomas based solely on the taxa within fecal samples has been limited. However, when 16S rRNA gene sequence data was combined with the results of a fecal immunochemical test (FIT), the ability to diagnose individuals with adenomas was improved relative to using the FIT results alone (12).

A recent meta-analysis found that 16S rRNA gene sequences from members of *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 that 16S rRNA gene sequences from members of *Haemophilus*, *Methanosphaera*, *Prevotella*, and *Succinovibrio* were enriched in patients with adenomas 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 a subset of the microbial community can be used to diagnose the presence of carcinoma.

Results

Lower bacterial diversity is associated with higher 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 of the communities were associated with disease stage after controlling for study and variable region differences. In fecal samples, both evenness and diversity were significantly lower in successive disease severity categories (evenness P-value=0.025 and diversity P-value=0.043) [Figure 1]; there was no significant difference for richness (P-value=0.21). We next tested whether the lower value of 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 (95% Confidence Interval: 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 ORs are significantly greater than 1.0, it is doubtful that they are clinically meaningful.

Similar to our analysis of sequences obtained from fecal samples, we repeated the analysis using sequences obtained from colon tissue. There were no significant differences in richness, evenness, or diversity as disease severity progressed from control to adenoma to carcinoma (P-values > 0.05). We next analyzed the ORs, 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 were not significantly different from 1.0 for any measure (P-values > 0.05) [Figure S1 & Table S1]. This is likely due to the combination of a small effect size and the relatively small number of studies and size of

studies used in the analysis.

Disease progression is associated with changes in community structure. 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 feces 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 fecal samples, there were significant differences in the bacterial community structure of subjects with normal colons and those with adenomas (1 of 2 studies) and carcinomas (1 of 3 studies) [Table S2]. For studies that used matched samples, we did not observe any differences in bacterial community structures [Table S2]. Combined, these results indicate that there were 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 that had 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 fecal or tissue samples (Table S3 & S4). In contrast, 8 taxa had significant ORs for the presence of carcinomas using data from fecal 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 a higher 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 XI* in fecal samples 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 fecal 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. OR values were caluclated 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. To measure the ability of these taxa to classify individuals we instead 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 classifying individuals by their disease status. Using data from fecal 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*. However, these had median AUC values less than 0.588 ndicating their limited value as biomarkers when used individually. Likewise, in unmatched tissue samples the 3 taxa with significant OR taxa had AUC values that 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 significant OR for the presences of carcinomas, they do a poor job of classifying a subject's disease status when used individually.

Combined taxa model classifies subjects better than using individual taxa. Instead of attempting to classify subjects based on individual taxa, next we combined information from the individual taxa and evaluated the ability to classify a subject's disease status using Random Forest models. For data from fecal 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). When this approach was used to train models using data from each study, the most important taxa were *Ruminococcus* and *Clostridium XI* [Figure 4A]. 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 classifying based on the relative abundances of *Blautia* and *Weissella* individually (P-value < 0.037). Both *Dorea* and *Blautia* were the most important taxa in the tissue-based models [Figure 4B]. Pooling the information from the taxa with significant ORs resulted in models that outperformed classifications made using the same taxa individually.

Performance of models based on taxa relative abundance in full community is better than that of models 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 fecal or tissue samples were only marginally better than 0.5 for any study (median AUC=0.549 (range: 0.367 - 0.971)) [Figure 5A & S2A]. In contrast, the models for classifying subjects as having normal colons or having carcinomas using

data from fecal or tissue samples yielded AUC values meaningfully higher than 0.5 [Figure 5B & S2B-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 fecal samples, we found that the AUC for 6 of 7 studies were an average of 14.8% higher and AUC for the Flemer study was 0.54% lower when using the relative abundance data from all taxa relative to using the relative abundance of only the taxa with significant ORs. The overall improvement in performance was statistically significant (mean=12.61%, one-tailed paired T-test; P-value=0.005). Among the models trained using data from fecal samples, *Bacteroides* and *Lachnospiraceae* were the most common taxa in the top 10% mean decrease in accuracy across studies [Figure S3]. Using data from unmatched tissue samples to train classification models, we found that the AUC of studies was an average 19.11% higher when we used all of the taxa rather than the 3 taxa with significant ORs (one-tailed paired T-test; P-value=0.03). For the models trained using data from unmatched tissue samples, *Lachnospiraceae*, *Bacteroidaceae*, and *Ruminococcaceae* were the most common taxa in the top 10% mean decrease in accuracy across studies [Figure S4]. Although the models trained using those taxa with a significant OR perform well for classifying individuals with and without carcinomas, models trained using data from the full community perform better.

Performance of models based on OTU relative abundances are not significantly better than those based on taxa with significant ORs. The previous models were based on relative abundance data where sequences were classified to coarse taxonomic assignments (i.e. typically genus or family level). To determine whether model performance improved with finer 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)). However, they did a good job of differentiating

between subjects with normal colons and those with carcinomas (median AUC: 0.71 (0.50-0.904)). 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) [Figure 4]. 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. This result was surprising as it indicated that a finer scale classification of the sequences 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 6 & S5]. As might be expected, the difference between the performance of the modelling approaches appeared to vary with the size of the training cohort (R^2 = 0.66) [Figure 6]. 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

We performed a meta-analysis to identify and validate microbiome-base biomarkers that could be used to classify individuals as having normal colons or colonic tumors using fecal or tissue samples. To our surprise, Random Forest classification models constructed to differentiate individuals with normal colons from those with carcinomas using a subset of the community performed well relative to models constructed using the full communities. When we applied the models trained on each dataset to the other datasets in our study, we found that the models trained using the subset of the communities performed better than those using the full communities. These models were trained using data in which sequences were assigned to bacterial taxa using a classifier that typically assigned sequences to the family or genus level. When we attempted to improve the specificity of the classification by using an OTU-based approach the resulting models performed as well as those constructed using coarse taxonomic assignments. These results are significant because they strengthen the growing literature indicating a role of the microbiome in tumorigenesis (9) and as a potential tool as a non-invasive diagnostic and for assessing risk of disease and recurrence (12, 30).

These results suggested that fine scale classification of sequences into OTUs does not improve our classification models. This has been suggested in previous literature where shotgun metagenomic data did not perform better than 16S rRNA gene sequencing data in classifying individuals with normal colons and those with carcinomas (31). We hypothesize that fine scale classification may not result in better classification because distribution of microbiota between individuals is patchy. In contrast, models using coarser taxonomic assignments will pool the fine scale diversity, resulting in less patchiness and better classification. Furthermore, the ability of models trained using a subset of the community to outperform those using the full community when testing the models on the other datasets may also be a product of the patchiness of the human-associated microbiota. The models

based on the 8 taxa that had significant ORs used taxa that were found in every study and tended to have higher relative abundances. Similar to the OTU-based models, those models based on the full community taxonomy assignments were still sensitive to the patchy distribution of taxa. Regardless, it is encouraging that a collection of 8 taxa could reliably classify individuals as having carcinomas considering the differences in cohorts, DNA extraction procedures, regions of the 16S rRNA gene, and sequencing methods.

When used separately to classify individuals with carcinomas, the taxa with significant ORs could not reliably classify individuals [Figure 3]. This result further supports the hypothesis that carcinoma-associated microbiota have a patchy distribution. Two individuals may have had the same classification, based on the relative abundance of different populations within this group of 8 taxa. Although these results only reflect associations with disease, it is tempting to hypothesize that the patchiness represents distinct mechanisms of exacerbating tumorigenesis or that multiple taxa have the same mechanism of exacerbating tumorigenesis. For example, strains of *Escherichia coli* and *Fusobacterium nucleatum* have been shown to worsen inflammation in mouse models of tumorigenesis (5, 6, 21). In contrast to the patchiness of the taxa that were positively associated with carcinomas, potentially beneficial taxa had a more consistent association [Figure 6]. This result was particularly interesting because members of these taxa (i.e. *Ruminococcus* and *Clostridium XI* in stool and *Dorea* and *Blautia* in tissue) are thought to be beneficial due to their involvement in production of anti-inflammatory short chain fatty acids (32–34).

All of the adenoma classification models performed poorly and is not inconsistent with previous studies (27, 30). However, the classification results are at odds with results of the multitarget microbiota test (MMT) from Baxter, *et al.* (12) who observed an AUC of 0.755 when applied to individuals with adenomas. There are two major differences between the models generated in this meta-analysis and that analysis. The MMT attempted to classify individuals as having a normal colon or having colonic

lesions (i.e. adenomas or carcinomas) and not adenomas alone. Further, the MMT incorporated fecal immunoglobulin test (FIT) data while our models only used 16S rRNA gene sequencing data. Because FIT data were not available for the other studies in our meta-analysis, it was not possible to validate the MMT approach. The ability to differentiate between individuals with and without adenomas is an important problem since early detection of tumors is critical. However, it is possible that we might have been able to detect differences in the bacterial community if individuals with non-advanced and advanced adenomas were separated. This is a clinically relevant distinction since advanced adenomas are at highest risk of progressing to a carcinoma. The initial changes of the microbiota during tumorigenesis could be focal to where the initial adenoma develops and would not be easily assessed using fecal samples from an individual with non-advanced adenomas. Unfortunately, distinguishing between individuals with advanced and non-advanced adenomas was not possible in our meta-analysis since the studies did not provide the clinical data needed to make that distinction.

Stool samples represent a non-invasive approach to assessing the structure of the gut microbiota and are potentially useful for diagnosing individuals as having colonic tumors. However, they do not reflect the structure of the mucosal microbiota (35). Regardless, the taxa that were the most important in the stool-based models overlapped with those from the models trained using the data from unmatched and matched colon tissue samples [Figure S3]. Mucosal biopsies are preferred for focused mechanistic studies and have offered researchers the opportunity to sample healthy and diseased tissue from the same individuals (i.e. matched) using each individual as their own control or in a cross sectional design (i.e. unmatched). Because obtaining these samples is invasive, carries risks to the individual, and is expensive, studies investigating the structure of the mucosal microbiota generally have a limited number of participants. Thus, it was not surprising that tissue-based studies did not provide clearer associations between the mucosal microbiota and the presence of tumors. Interestingly, Fusobacterium, which

has received increased recent attention for its potential role in tumorigenesis (6) was not consistently identified across the studies in our meta-analysis. This could be due to the relatively small number of individuals in the limited number of studies. The classification models trained using the tissue-based data performed well when tested with the training data (Figure S4), but performed poorly when tested on the other tissue-associated datasets (Figure S5). Disturbingly, taxa that are commonly associated with reagent contamination (e.g. *Novosphingobium*, *Acidobacteria Gp2*, *Sphingomonas*, etc.) were detected within the tissue datasets. Such contamination is common in studies where there is relatively low bacterial biomass (36). The lack of replication among the tissue-based biomarkers may be a product of the relatively small number of studies and individuals per study and possible reagent contamination.

Among our stool samples, we failed to identify several notable populations that are commonly associated with carcinomas including an enterotoxigenic strain of *Bacteroides fragilis* (ETBF) and *Streptococcus gallolyticus* subsp. *gallolyticus* (22, 24). ETBF have been found in tumors in the proximal colon where they tend to form biofilms (20, 37). Considering DNA from bacteria that are more prevalent in the proximal colon may be degraded by the time it leaves the body, it is not surprising that we failed to identify a significant OR for Bacteroides with carcinomas. In addition, since our approach could only classify sequences to the genus level and there are likely multiple Bacteroides populations in the colon, it is possible that sequences from ETBF and non-oncogenic Bacteroides were pooled. This would then reduce the OR between Bacterioides and whether an individual had carcinomas. It is also necessary to distinguish between populations that are biomarkers for a disease and those that are known to cause disease. The former may also have a causative role in the disease. Although the latter have been shown to have a causative role, they may appear at low relative abundance, be found in specific locations, or may have a highly patchy distribution among affected individuals.

Meta-analyses are a useful tool in microbiome research because they can demonstrate whether a result can be replicated and facilitate new discoveries by pooling multiple independent investigations. There have been several meta-analyses similar to this study that have sought biomarkers for obesity (38–40), inflammatory bowel disease (39), and colorectal cancer (28). Considering microbiome research is particularly prone to hype and overgeneralization of results, these analyses are critical. For example, previous meta-analyses have demonstrated that there are no clear fecal biomarkers for obesity (39, 40). Such meta-analyses are difficult to perform because the underlying 16S rRNA gene sequence data are not publicly available, metadata are missing, incomplete, or vague, sequence data are of poor quality or derived by non-standard approaches, and the original studies were significantly underpowered. Reluctance to publish negative results (i.e. the "file drawer effect") is also likely to skew our understanding of the relationship between microbiota and disease. Better attention to these specifics will increase the reproducibility and replicability of microbiota studies and make it easier to perform these crucial meta-analyses. Moving forward, meta-analyses will be important tools to help aggregate and find commonalities across studies when investigating the microbiota in the context of a specific disease (28, 38–40).

Our meta-analysis suggests a strong association between the gut microbiota and colon tumorigenesis. By aggregating the results from studies that sequenced the 16S rRNA gene from fecal and tissue samples, we are able to provide evidence supporting the use of microbial biomarkers to diagnose the presence of colonic tumors. Further development of microbial biomarkers should focus on including other biomarkers (e.g. FIT), better categorizing of people with adenomas, and expanding datasets to include larger numbers of individuals. Based on prior research into the physiology of the biomarkers we identified, it is likely that they have a causative role in tumorigenesis. Their patchy distribution across individuals suggests that there are either multiple mechanisms causing disease or a single mechanism (e.g. inflammation) that can be mediated by multiple, diverse bacteria.

Methods

Datasets. The studies used for this meta-analysis were identified through the review articles written by Keku, et al. (41) and Vogtmann, et al. (42). Additional studies, not mentioned in those reviews were obtained based on the authors' knowledge of the literature. Studies were included that used tissue or feces as their sample source for 454 or Illumina 16S rRNA gene sequencing. Some studies (N = 12) were excluded because they did not have publicly available sequences, did not use 454 or Illumina machines, or did not have metadata in which the authors were able to share. We were able to obtain sequence data and metadata from the following studies: 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.* (43), Lu, *et al.* (26), Sanapareddy, *et al.* (25), Wang, *et al.* (14), Weir, *et al.* (23), and Zeller, *et al.* (16). The Zackular (44) study was excluded because their 90 individuals were contained within the larger Baxter study (12). The Kostic study was excluded because after we processed the sequences, all of the case samples had 100 or fewer sequences. The final analysis included 14 studies (Tables 1 and 2). 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). After curating the sequences, 1737 stool samples and 492 tissue samples remained in the analysis [Tables 1 and 2].

Sequence Processing. Raw sequence data and metadata were primarily obtained from the Sequence Read Archive (SRA) and dbGaP. Other sequence and metadata were obtained directly from the authors (n = 4, (17, 23, 25, 27)). Each dataset was processed separately using mothur (v1.39.3) (45) using the default quality filtering methods for both 454 and Illumina sequence data. If it was not possible to use the defaults because the sequences were trimmed too much, then the stated quality cut-offs from the original study

were used. Chimeric sequences were identified and removed using VSEARCH (46). The curated sequences were assigned to OTUs at 97% similarity using the OptiClust algorithm (47) and classified to the deepest taxonomic level that had 80% support using the naïve Bayesian classifier trained on the RDP taxonomy outline (version 14, (48)).

Community analysis. We calculated alpha diversity metrics (i.e. OTU richness, evenness, and Shannon diversity) for each sample. Within each dataset, we ensured that the data followed a normal distribution using power transformations. Using the transformed data, we tested the hypothesis that individuals with normal colons, adenomas, and carcinomas had significantly different alpha diversity metrics using linear mixed-effect models. We also calculated the OR for each study and metric by considering any value above the median alpha diversity value to be positive. We measured the dissimilarity between individuals by calculating the pairwise Bray-Curtis index and used PERMANOVA (49) to test whether individuals with normal colons were significantly different from those with adenomas or carcinomas. Finally, after binning sequences into the deepest taxa that the naïve Bayesian classifier could calssify the sequences, we quantified the ORs for individuals having an adenoma or carcinoma and corrected for multiple comparisons using the Benjamini-Hochberg method (50). Again, for each taxon, if the relative abundance was greater than the median relative abundance for that taxon in the study, the individual was considered to be positive.

Random Forest classification analysis. To classify individuals as having normal colons or tumors, we built Random Forest classification models for each dataset and comparison using taxa with significant ORs (after multiple comparison correction), all taxa, or OTUs. Because no taxa were identified as having a significant OR associated with adenomas using stool samples or tissue samples, classification models based on OR data were not constructed to classify individuals as having normal colons or adenomas. Within the training dataset, 10-fold cross validation (5-fold cross validation for small datasets)

was used to build a model that was then evaluated on the testing set. For the models constructed using the taxa with significant ORs, the default mtry setting was used to train the model and this model was tested on the other datasets in the meta-analysis. The reported AUC values are the AUCs for the application of the model on the test sets. For the OTU-based models, the dataset was split into training (80% of samples) and testing (20%) sets and 10-fold cross validation (5-fold cross validation for small datasets) on the training set was used to generate the model for the testing set. The original 80/20 split and fitting was repeated 100 times and the average AUC from these 100 repeats was reported. The Mean Decrease in Accuracy (MDA), a measure of the importance of each taxon to the overall model was used to rank the taxa used in each model. For all models, the default setting used was $\sqrt{p}$, where p is all the variables used in the respective model. Normally, $\sqrt{p}$ has been found to be what is chosen as the ideal mtry after model tuneing (51).

Statistical Analysis. All statistical analysis after sequence processing utilized the R (v3.4.3) software package (52). For OTU richness, evenness, and Shannon diversity analysis, values were power transformed using the rcompanion (v1.11.1) package (53) and then Z-score normalized using the car (v2.1.6) package (54). Testing for OTU richness, evenness, and Shannon diversity differences utilized linear mixed-effect models created using the lme4 (v1.1.15) package (55) 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 (56, 57) 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 (58). Random Forest models were built using both the caret (v6.0.78) and randomForest (v4.6.12) packages (59, 60). All figures were created using both ggplot2 (v2.2.1) and gridExtra (v2.3) packages (61, 62).

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.

Acknowledgements

The authors would like to thank all the study participants who were a part of each of the individual studies analyzed. We would also like to thank each of the study authors for making their sequencing reads and metadata available for use. Finally, we would like to thank the members of the Schloss lab for their valuable feedback and proofreading during the formulation of this manuscript.

References

1. **Siegel, R. L.**, **K. D. Miller**, and **A. Jemal**. 2016. Cancer statistics, 2016. CA: a cancer journal for clinicians **66**:7–30.

2. **Flynn, K. J.**, **N. T. Baxter**, and **P. D. Schloss**. 2016. Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. mSphere **1**.

3. **Goodwin, A. C.**, **C. E. Destefano Shields**, **S. Wu**, **D. L. Huso**, **X. Wu**, **T. R. Murray-Stewart**, **A. Hacker-Prietz**, **S. Rabizadeh**, **P. M. Woster**, **C. L. Sears**, and **R. A. Casero**. 2011. Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. Proceedings of the National Academy of Sciences of the United States of America **108**:15354–15359.

4. **Abed, J.**, **J. E. M. Emgård**, **G. Zamir**, **M. Faroja**, **G. Almogy**, **A. Grenov**, **A. Sol**, **R. Naor**, **E. Pikarsky**, **K. A. Atlan**, **A. Mellul**, **S. Chaushu**, **A. L. Manson**, **A. M. Earl**, **N. Ou**, **C. A. Brennan**, **W. S. Garrett**, and **G. Bachrach**. 2016. Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. Cell Host & Microbe **20**:215–225.

5. **Arthur, J. C.**, **E. Perez-Chanona**, **M. Mühlbauer**, **S. Tomkovich**, **J. M. Uronis**, **T.-J. Fan**, **B. J. Campbell**, **T. Abujamel**, **B. Dogan**, **A. B. Rogers**, **J. M. Rhodes**, **A. Stintzi**, **K. W. Simpson**, **J. J. Hansen**, **T. O. Keku**, **A. A. Fodor**, and **C. Jobin**. 2012. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science (New York, N.Y.) **338**:120–123.

6. **Kostic, A. D.**, **E. Chun**, **L. Robertson**, **J. N. Glickman**, **C. A. Gallini**, **M. Michaud**, **T. E. Clancy**, **D. C. Chung**, **P. Lochhead**, **G. L. Hold**, **E. M. El-Omar**, **D. Brenner**, **C. S. Fuchs**, **M. Meyerson**, and **W. S. Garrett**. 2013. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host &

Microbe **14**:207–215.

7. **Wu, S.**, **K.-J. Rhee**, **E. Albesiano**, **S. Rabizadeh**, **X. Wu**, **H.-R. Yen**, **D. L. Huso**, **F. L. Brancati**, **E. Wick**, **F. McAllister**, **F. Housseau**, **D. M. Pardoll**, and **C. L. Sears**. 2009. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nature Medicine **15**:1016–1022.

8. **Zackular, J. P.**, **N. T. Baxter**, **G. Y. Chen**, and **P. D. Schloss**. 2016. Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. mSphere **1**.

9. **Zackular, J. P.**, **N. T. Baxter**, **K. D. Iverson**, **W. D. Sadler**, **J. F. Petrosino**, **G. Y. Chen**, and **P. D. Schloss**. 2013. The gut microbiome modulates colon tumorigenesis. mBio **4**:e00692–00613.

10. **Baxter, N. T.**, **J. P. Zackular**, **G. Y. Chen**, and **P. D. Schloss**. 2014. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. Microbiome **2**:20.

11. **Ahn, J.**, **R. Sinha**, **Z. Pei**, **C. Dominianni**, **J. Wu**, **J. Shi**, **J. J. Goedert**, **R. B. Hayes**, and **L. Yang**. 2013. Human gut microbiome and risk for colorectal cancer. Journal of the National Cancer Institute **105**:1907–1911.

12. **Baxter, N. T.**, **M. T. Ruffin**, **M. A. M. Rogers**, and **P. D. Schloss**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**:37.

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

14. **Wang, T.**, **G. Cai**, **Y. Qiu**, **N. Fei**, **M. Zhang**, **X. Pang**, **W. Jia**, **S. Cai**, and **L. Zhao**. 2012. Structural segregation of gut microbiota between colorectal cancer patients and

healthy volunteers. The ISME journal **6**:320–329.

15. **Burns, M. B.**, **J. Lynch**, **T. K. Starr**, **D. Knights**, and **R. Blekhman**. 2015. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome Medicine **7**:55.

16. **Zeller, G.**, **J. Tap**, **A. Y. Voigt**, **S. Sunagawa**, **J. R. Kultima**, **P. I. Costea**, **A. Amiot**, **J. Böhm**, **F. Brunetti**, **N. Habermann**, **R. Hercog**, **M. Koch**, **A. Luciani**, **D. R. Mende**, **M. A. Schneider**, **P. Schrotz-King**, **C. Tournigand**, **J. Tran Van Nhieu**, **T. Yamada**, **J. Zimmermann**, **V. Benes**, **M. Kloor**, **C. M. Ulrich**, **M. von Knebel Doeberitz**, **I. Sobhani**, and **P. Bork**. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology **10**:766.

17. **Flemer, B.**, **D. B. Lynch**, **J. M. R. Brown**, **I. B. Jeffery**, **F. J. Ryan**, **M. J. Claesson**, **M. O'Riordain**, **F. Shanahan**, and **P. W. O'Toole**. 2017. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut **66**:633–643.

18. **García, A. Z. G.** 2012. Factors influencing colorectal cancer screening participation. Gastroenterology Research and Practice. Hindawi Limited **2012**:1–8.

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

20. **Dejea, C. M.**, **E. C. Wick**, **E. M. Hechenbleikner**, **J. R. White**, **J. L. Mark Welch**, **B. J. Rossetti**, **S. N. Peterson**, **E. C. Snesrud**, **G. G. Borisy**, **M. Lazarev**, **E. Stein**, **J. Vadivelu**, **A. C. Roslani**, **A. A. Malik**, **J. W. Wanyiri**, **K. L. Goh**, **I. Thevambiga**, **K. Fu**, **F. Wan**, **N. Llosa**, **F. Housseau**, **K. Romans**, **X. Wu**, **F. M. McAllister**, **S. Wu**, **B. Vogelstein**, **K. W. Kinzler**, **D. M. Pardoll**, and **C. L. Sears**. 2014. Microbiota organization is a distinct feature of proximal colorectal cancers. Proceedings of the National Academy

of Sciences of the United States of America **111**:18321–18326.

21. **Arthur, J. C.**, **R. Z. Gharaibeh**, **M. Mühlbauer**, **E. Perez-Chanona**, **J. M. Uronis**, **J. McCafferty**, **A. A. Fodor**, and **C. Jobin**. 2014. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. Nature Communications. Springer Nature **5**:4724.

22. **Aymeric, L.**, **F. Donnadieu**, **C. Mulet**, **L. du Merle**, **G. Nigro**, **A. Saffarian**, **M. Bérard**, **C. Poyart**, **S. Robine**, **B. Regnault**, **P. Trieu-Cuot**, **P. J. Sansonetti**, and **S. Dramsi**. 2017. Colorectal cancer specific conditions promoteStreptococcus gallolyticusgut colonization. Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences **115**:E283–E291.

23. **Weir, T. L.**, **D. K. Manter**, **A. M. Sheflin**, **B. A. Barnett**, **A. L. Heuberger**, and **E. P. Ryan**. 2013. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. PloS One **8**:e70803.

24. **Boleij, A.**, **E. M. Hechenbleikner**, **A. C. Goodwin**, **R. Badani**, **E. M. Stein**, **M. G. Lazarev**, **B. Ellis**, **K. C. Carroll**, **E. Albesiano**, **E. C. Wick**, **E. A. Platz**, **D. M. Pardoll**, and **C. L. Sears**. 2014. The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. Clinical Infectious Diseases. Oxford University Press (OUP) **60**:208–215.

25. **Sanapareddy, N.**, **R. M. Legge**, **B. Jovov**, **A. McCoy**, **L. Burcal**, **F. Araujo-Perez**, **T. A. Randall**, **J. Galanko**, **A. Benson**, **R. S. Sandler**, **J. F. Rawls**, **Z. Abdo**, **A. A. Fodor**, and **T. O. Keku**. 2012. Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans. The ISME journal **6**:1858–1868.

26. **Lu, Y.**, **J. Chen**, **J. Zheng**, **G. Hu**, **J. Wang**, **C. Huang**, **L. Lou**, **X. Wang**, and **Y. Zeng**. 2016. Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas.

Scientific Reports **6**:26337.

27. **Hale, V. L.**, **J. Chen**, **S. Johnson**, **S. C. Harrington**, **T. C. Yab**, **T. C. Smyrk**, **H. Nelson**, **L. A. Boardman**, **B. R. Druliner**, **T. R. Levin**, **D. K. Rex**, **D. J. Ahnen**, **P. Lance**, **D. A. Ahlquist**, and **N. Chia**. 2017. 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 **26**:85–94.

28. **Shah, M. S.**, **T. Z. DeSantis**, **T. Weinmaier**, **P. J. McMurdie**, **J. L. Cope**, **A. Altrichter**, **J.-M. Yamal**, and **E. B. Hollister**. 2017. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. Gut.

29. **Brim, H.**, **S. Yooseph**, **E. G. Zoetendal**, **E. Lee**, **M. Torralbo**, **A. O. Laiyemo**, **B. Shokrani**, **K. Nelson**, and **H. Ashktorab**. 2013. Microbiome analysis of stool samples from African Americans with colon polyps. PloS One **8**:e81352.

30. **Sze, M. A.**, **N. T. Baxter**, **M. T. Ruffin**, **M. A. M. Rogers**, and **P. D. Schloss**. 2017. Normalization of the microbiota in patients after treatment for colonic lesions. Microbiome. Springer Nature **5**.

31. **Hannigan, G. D.**, **M. B. Duhaime**, **M. T. Ruffin**, **C. C. Koumpouras**, and **P. D. Schloss**. 2017. Diagnostic potential & the interactive dynamics of the colorectal cancer virome. Cold Spring Harbor Laboratory.

32. **Venkataraman, A.**, **J. R. Sieber**, **A. W. Schmidt**, **C. Waldron**, **K. R. Theis**, and **T. M. Schmidt**. 2016. Variable responses of human microbiomes to dietary supplementation with resistant starch. Microbiome. Springer Nature **4**.

33. **Herrmann, E.**, **W. Young**, **V. Reichert-Grimm**, **S. Weis**, **C. Riedel**, **D. Rosendale**, **H. Stoklosinski**, **M. Hunt**, and **M. Egert**. 2018. In vivo assessment of resistant starch

581 degradation by the caecal microbiota of mice using RNA-based stable isotope probingA
582 proof-of-principle study. Nutrients. MDPI AG **10**:179.

583 34. **Reichardt, N.**, **M. Vollmer**, **G. Holtrop**, **F. M. Farquharson**, **D. Wefers**, **M. Bunzel**,
584 **S. H. Duncan**, **J. E. Drew**, **L. M. Williams**, **G. Milligan**, **T. Preston**, **D. Morrison**, **H. J.**
585 **Flint**, and **P. Louis**. 2017. Specific substrate-driven changes in human faecal microbiota
586 composition contrast with functional redundancy in short-chain fatty acid production. The
587 ISME Journal. Springer Nature **12**:610–622.

588 35. **Flynn, K. J.**, **M. T. Ruffin**, **D. K. Turgeon**, and **P. D. Schloss**. 2017. Spatial variation
589 of the native colon microbiota in healthy adults. Cold Spring Harbor Laboratory.

590 36. **Salter, S. J.**, **M. J. Cox**, **E. M. Turek**, **S. T. Calus**, **W. O. Cookson**, **M. F. Moffatt**,
591 **P. Turner**, **J. Parkhill**, **N. J. Loman**, and **A. W. Walker**. 2014. Reagent and laboratory
592 contamination can critically impact sequence-based microbiome analyses. BMC Biology.
593 Springer Nature **12**.

594 37. **Purcell, R. V.**, **J. Pearson**, **A. Aitchison**, **L. Dixon**, **F. A. Frizelle**, and **J. I. Keenan**.
595 2017. Colonization with enterotoxigenic bacteroides fragilis is associated with early-stage
596 colorectal neoplasia. PLOS ONE. Public Library of Science (PLoS) **12**:e0171602.

597 38. **Sze, M. A.**, and **P. D. Schloss**. 2016. Looking for a signal in the noise: Revisiting
598 obesity and the microbiome. mBio. American Society for Microbiology **7**:e01018–16.

599 39. **Walters, W. A.**, **Z. Xu**, and **R. Knight**. 2014. Meta-analyses of human gut microbes
600 associated with obesity and IBD. FEBS Letters. Wiley-Blackwell **588**:4223–4233.

601 40. **Finucane, M. M.**, **T. J. Sharpton**, **T. J. Laurent**, and **K. S. Pollard**. 2014. A taxonomic
602 signature of obesity in the microbiome? Getting to the guts of the matter. PLoS ONE.

603 Public Library of Science (PLoS) **9**:e84689.

604 41. **Keku, T. O.**, **S. Dulal**, **A. Deveaux**, **B. Jovov**, and **X. Han**. 2015. The gastrointestinal
605 microbiota and colorectal cancer. American Journal of Physiology - Gastrointestinal and
606 Liver Physiology **308**:G351–G363.

607 42. **Vogtmann, E.**, and **J. J. Goedert**. 2016. Epidemiologic studies of the human
608 microbiome and cancer. British Journal of Cancer **114**:237–242.

609 43. **Kostic, A. D.**, **D. Gevers**, **C. S. Pedamallu**, **M. Michaud**, **F. Duke**, **A. M. Earl**, **A. I.**
610 **Ojesina**, **J. Jung**, **A. J. Bass**, **J. Tabernero**, **J. Baselga**, **C. Liu**, **R. A. Shivdasani**, **S.**
611 **Ogino**, **B. W. Birren**, **C. Huttenhower**, **W. S. Garrett**, and **M. Meyerson**. 2012. Genomic
612 analysis identifies association of Fusobacterium with colorectal carcinoma. Genome
613 Research **22**:292–298.

614 44. **Zackular, J. P.**, **M. A. M. Rogers**, **M. T. Ruffin**, and **P. D. Schloss**. 2014. The human
615 gut microbiome as a screening tool for colorectal cancer. Cancer Prevention Research
616 (Philadelphia, Pa.) **7**:1112–1121.

617 45. **Schloss, P. D.**, **S. L. Westcott**, **T. Ryabin**, **J. R. Hall**, **M. Hartmann**, **E. B. Hollister**,
618 **R. A. Lesniewski**, **B. B. Oakley**, **D. H. Parks**, **C. J. Robinson**, **J. W. Sahl**, **B. Stres**, **G.**
619 **G. Thallinger**, **D. J. Van Horn**, and **C. F. Weber**. 2009. Introducing mothur: Open-Source,
620 Platform-Independent, Community-Supported Software for Describing and Comparing
621 Microbial Communities. Appl.Environ.Microbiol. **75**:7537–7541.

622 46. **Rognes, T.**, **T. Flouri**, **B. Nichols**, **C. Quince**, and **F. Mahé**. 2016. VSEARCH: A
623 versatile open source tool for metagenomics. PeerJ **4**:e2584.

624 47. **Westcott, S. L.**, and **P. D. Schloss**. 2017. OptiClust, an Improved Method for

625 Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere **2**.

626 48. **Wang, Q.**, **G. M. Garrity**, **J. M. Tiedje**, and **J. R. Cole**. 2007. Naive bayesian classifier
627 for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and
628 Environmental Microbiology. American Society for Microbiology **73**:5261–5267.

629 49. **Anderson, M. J.**, and **D. C. I. Walsh**. 2013. PERMANOVA, ANOSIM, and the mantel
630 test in the face of heterogeneous dispersions: What null hypothesis are you testing?
631 Ecological Monographs. Wiley-Blackwell **83**:557–574.

632 50. **Benjamini, Y.**, and **Y. Hochberg**. 1995. Controlling the false discovery rate: A
633 practical and powerful approach to multiple testing. Journal of the Royal Statistical Society.
634 Series B (Methodological) **57**:289–300.

635 51. **Breiman, L.** 2001. Machine Learning. Springer Nature **45**:5–32.

636 52. **R Core Team**. 2017. R: A language and environment for statistical computing. R
637 Foundation for Statistical Computing, Vienna, Austria.

638 53. **Mangiafico, S.** 2017. Rcompanion: Functions to support extension education program
639 evaluation.

640 54. **Fox, J.**, and **S. Weisberg**. 2011. An R companion to applied regressionSecond. Sage,
641 Thousand Oaks CA.

642 55. **Bates, D.**, **M. Mächler**, **B. Bolker**, and **S. Walker**. 2015. Fitting linear mixed-effects
643 models using lme4. Journal of Statistical Software **67**:1–48.

644 56. **Telmo Nunes, M. S. with contributions from**, **C. Heuer**, **J. Marshall**, **J. Sanchez**,
645 **R. Thornton**, **J. Reiczigel**, **J. Robison-Cox**, **P. Sebastiani**, **P. Solymos**, **K. Yoshida**, **G.**
646 **Jones**, **S. Pirikahu**, **S. Firestone**, and **R. Kyle.** 2017. EpiR: Tools for the analysis of

epidemiological data.

57. **Viechtbauer, W.** 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software **36**:1–48.

58. **Oksanen, J.**, **F. G. Blanchet**, **M. Friendly**, **R. Kindt**, **P. Legendre**, **D. McGlinn**, **P. R. Minchin**, **R. B. O'Hara**, **G. L. Simpson**, **P. Solymos**, **M. H. H. Stevens**, **E. Szoecs**, and **H. Wagner**. 2017. Vegan: Community ecology package.

59. **Jed Wing, M. K. C. from**, **S. Weston**, **A. Williams**, **C. Keefer**, **A. Engelhardt**, **T. Cooper**, **Z. Mayer**, **B. Kenkel**, **the R Core Team**, **M. Benesty**, **R. Lescarbeau**, **A. Ziem**, **L. Scrucca**, **Y. Tang**, **C. Candan**, and **T. Hunt.** 2017. Caret: Classification and regression training.

60. **Liaw, A.**, and **M. Wiener**. 2002. Classification and regression by randomForest. R News **2**:18–22.

61. **Wickham, H.** 2009. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

62. **Auguie, B.** 2017. GridExtra: Miscellaneous functions for "grid" graphics.

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: Most Important Members in Significant OR Taxa Carcinoma Models. A) Common taxa in the top 10 percent for carcinoma Random Forest stool-based models. B) Common taxa in the top 10 percent for carcinoma Random Forest unmatched tissue-based models. Blue represents less important and red represents more important to the Random Forest Model. White means that particular taxa was not in the top 10%.

Figure 5: 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 6: 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

689 all genera or significant OR taxa only. The black line represents the AUC at 0.5. The red
690 lines represent the median AUC of all test AUCs for a specific study.

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: 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 all OTU-based models. Blue represents less important and red represents more important to the Random Forest Model. White means that particular taxa was not in the top 10%.

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. Blue represents less important and red represents more important to the Random Forest Model. White means that particular taxa was not in the top 10%.

Figure S5: 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.