Investigating the Microbiota and Colorectal Cancer: The Importance of Community

Marc A Sze[1] and Patrick D Schloss[1†]

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

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

Co-author e-mails:

- marcsze@med.umich.edu

Abstract

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

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

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

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

Keywords

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

Background

Colorectal cancer (CRC) is a growing world-wide health problem [1] in which the microbiota has been purported to play an active role in disease pathogenesis [2]. Numerous studies have shown the importance of both individual microbes [3–7] and the overall community [8–10] in polyp formation using mouse models of CRC. There have also been numerous case/control studies investigating the microbiota in the formation of both adenoma and carcinoma. Recently, a meta-analysis was published investigating whether specific biomarkers could be consistently identified using multiple data sets [11]. Many of the studies, along with the previous meta-analysis, focus on identifying biomarkers or individual microbes but do not critically investigate how the community changes in CRC.

Targeting the identification of biomarkers within stool seems logical since the current gold standard for diagnosis, a colonoscopy, can be time-consuming and is not risk free. Stool offers an easy and cost-effective way in which to stratify CRC risk within the population. Although stool represents an easy and less invasive way to assess risk, it is not clear how reflective this sample is to what is actually happening on the adenoma or carcinoma. Some studies have begun to try and assess this in health and disease but are limited by their sample size [12,13]. Since the stool possibly represents an agregate community of the overall gastrointestinal (GI) tract, sampling the adenoma or carinoma directly could more accurately identify important microbes which could be drastically different then what is identified in stool. Sampling the tissue directly may provide clearer answers but it is not without problems. Due to bowel prep the communities left for sampling may not be reflective of the resident microbiota, but rather a collection of what is able to keep adhered to the mucosa. Additionally, these samples would contain much more host DNA then microbial DNA, potentially limiting the types of analysis that can be done. It has also been well published that low biomass samples can be very difficult to work with and results can end up being study dependent due to the randomness of contamination [14]. Due to these many differences, one question our meta-analysis aims to answer is whether it is possible to find consistent patterns that emerge across studies regardless of whether they used stool or tissue samples.

This intense focus on identifying biomarkers has identified mostly mouth-associated microbes as potential CRC-associated microbes in the GI tract [15–17]. The main bacteria of interest arising from this set of microbes has been those within the *Fusobacterium* genus. Yet, the question remains as to whether or not this is indeed the most important genera to be focusing on, since many microbiota-based studies typically have identified a collection of oral microbes rather than single species from a single genera [16,17]. Based on this discrepency, the second question we want to answer with this meta-analysis is if there is one dominant CRC-associated genera that can be identified across studies.

The heavy concentration on the identification of biomarkers has mostly dominated the study of CRC and has had an unattended consequence of reducing the focus on changes that occur within the underlying resident community. This has been borne out by the majority of previous studies, within stool, tissue, and the only meta-analysis investigating this area, that have focused predominately on biomarker identification. Since it has not been investigated in detail the final question that this meta-analysis aims to answer is whether there are consistent detectable community differences as disease severity increases.

In comparison to the previous meta-analysis, this study significantly increases the total stool samples investigated, examines differences between stool and tissue microbiota in the context of CRC, and takes a more community centric approach rather than a biomarker focused approach to investigating commonalities across study for the microbiota and CRC severity. Importantly, this community centric approach could provide valuable insights into the importance of accounting for the community in CRC disease not previously provided by earlier meta-analysis studies [11].

Using both feces (n = 1737) and tissue (492 samples from 350 individuals) totalling over 2229 total samples across 14 studies [12,16–28] [Table 1 & 2], we expand both the breadth and scope of the previous meta-analysis to investigate whether the bacterial community is an important risk factor for both adenoma and carcinoma. To accomplish this we first assessed whether the diversity changes throughout disease (control to adenoma to carcinoma) and if it results in an increased relative risk (RR) for adenoma or carcinoma. We then assessed how common CRC-associated genera (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, or *Porphyromonas*) affect the relative risk of adenoma or carcinoma. Next, using Random Forest models, we analyzed whether the full community or only the CRC-associated genera resulted in better model classification based on the area under the curve (AUC). We observed that the community changes as disease severity worsens and that this community is important for disease classification. However, since the changes in community were subtle for adenoma we also examined what effect and sample size the studies that were used were adequately powered for. Although we analyzed data sets which sampled large numbers of individuals, our results indicate the individual studies were underpowered for detecting effect size differences of 10% or below between the case and control groups.

Results

Lower Community Diversity is Associated with Increased RR of Carcinomas: Using the combined data set we first assessed whether there were any broad scale community differences that could be detected as disease severity worsened. Using power transformed and Z-score normalized α-diversity metrics, both evenness and Shannon diversity in feces, but not tissue, were lower in those with carcinoma [Figure 1]. Using linear mixed-effect models to control for study and variable region, there was a significant decrease from control to adenoma to carcinoma for both evenness (P-value = 0.025) and Shannon diversity (P-value = 0.043). However, in tissue, this effect was not observed when resampling of the same individual was also controlled for (P-value > 0.05). We next tested whether these detectable differences in the community resulted in a significant increase in RR. Within fecal samples, a decrease in Shannon diversity and evenness resulted in a significantly increased RR for carcinoma (P-value = 0.0047 and 0.00034, respectively) [Figure 2]. Although these values were significant, the effect size was relatively small for both metrics (Shannon RR = 1.33 (1.09 - 1.62) and evenness RR = 1.36 (1.15 - 1.61)) [Figure 2]. Only a decrease in evenness had an increased RR for adenoma (P-value = 0.032) [Figure 2A & S1] but this effect size was even smaller than what was observed for carcinoma (RR = 1.16 (1.01 - 1.34)). Interestingly, for both adenoma and carcinoma there was no increase in RR within tissue samples for any alpha diversity metric investigated [Table S1-S3].

Using the Bray-Curtis distance metric, there was a significant difference across studies in the bacterial community of fecal samples between carcinoma and controls, but not adenoma and controls [Table S4 & S5]. For studies with unmatched tissue samples a similar trend was observed [Table S3 & S4] while studies with tissue samples from the same individual (matched) had no differences [Table S6 & S7].

Carcinoma-Associated Genera Minimally Impacts RR of Adenoma: Based on the small increase in RR using α-diversity metrics, we wanted to know if the presence of specific genera resulted in a higher RR for both stool and tissue. To investigate this we analyzed the classically associated CRC genera, *Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, and *Porphyromonas* for and increase in RR. The majority of CRC-associated genera for both feces and tissue had a significantly increased RR for carcinoma but not for adenoma [Figure 3]. In fecal samples the RR due to CRC associated genera was greater than either the RR assoicated with evenness or Shannon diversity [Figure 2 & 3]. Additionally, the RR of carcinoma continuously increased as individuals tested positive for more CRC associated genera [Figure 3B & 3D]. The RR effect size was greater for stool (RR range = 1.62 - 2.37) than for tissue (RR range = 1.21 - 1.81). This decrease may be explained by the fact that the tissue analysis included matched samples from the same individual.

There were two significant measures for increased RR of adenoma when investigating CRC-associated genera in stool: 1) Having a higher then median value of *Porphyromonas* (P-value = 0.023) and 2) whether samples were positive for three CRC associated genera (P-value = 0.022) [Figure 3A]. With tissue, there were three significant measures for an increased RR of adenoma: 1) being positive for one CRC-associated genera (P-value = 0.032), 2) being positive for two CRC associated genera (P-value = 0.008), and 3) being positive for four CRC associated genera (P-value = 0.039) [Figure 3C].

Using the Whole Community Instead of Only CRC-Associated Genera Increases Model AUC: We then tested whether the overall bacterial community was at all important to classifying disease or if the CRC-associated genera were sufficient alone. To test this we used two approaches. The first used genus level data and tested whether there were any differences in AUC when training on one study and testing on all the others when using either all genera present or only the CRC-associated genera. The second approach used OTU level data and tested whether there was a generalized decrease in the 10-fold cross validation (CV) model across studies using either all OTUs or only OTUs that taxonomically classified to CRC-associated genera.

The genus based models showed an AUC decrease in model classification on the training set for both stool and tissue studies [Figure S2-S3]. With respect to the test sets, comprised of genera data from other studies, both the all genera model and CRC-associated models had a similar ability to detect adenomas or carcinomas [Figure S4-S6]. The AUC for classification of adenomas was lower than carcionmas for both tissue and stool [Figure S4-S6]. Interestingly, the AUC for the classification of carcinoma was consistently lower for the tissue models than the stool models [Figure S4-S6].

The OTU based models for both fecal and tissue (matched and unmatched) samples, showed an AUC decrease when only OTUs from the CRC-associated genera are used versus the full community of OTUs [Figure 4 & 5]. This decrease is observed in both adenoma and carcinoma groups [Figure 4 & 5] with the largest difference in median AUC for stool being in carcinoma classification [Figure 4B] and for tissue being an adenoma [Figure 5A].

In stool the most common genera in the genus based models belonged mostly to resident genera [Figure 6A & B]. This included genera such as *Ruminococcus*, *Bacteroides*, and *Roseburia*. With respect to the CRC-associated genera *Fusobacterium* was the only genus present in adenoma while all four were present in carcinoma [Figure 6A & B]. However, none of these CRC-associated genera were present in the majority of studies. For the adenoma OTU models, OTUs that classified as *Ruminococcaceae* or *Roseburia* were present in the top 10 OTUs for the vast majority of studies [Figure 6C]. Although for the carcinoma OTU models *Ruminococcaceae* was in the top 10 for many studies, *Bacteroides* was present in the overwhelming majority of the carcinoma OTU stool models [Figure 6D].

Conversely to the stool models, both genera and OTU based models in tissue had the vast majority of their top 10 occur in a study specific manner [Figure S7]. *Fusobacterium* and *Fusobacteriaceae* show up more often in the top 10 for matched tissue samples but are not present in the top 10 or much lower for unmatched tissue [Figure S7B-C & S7E-F]. There appears to be very little overlap in the top 10 most important variables between stool and tissue for both adenoma and carcinoma [Figure 6 & S7].

A Majority of Studies are Underpowered for Detecting Small Effect Size Differences: Based on the previous observations we then assessed whether the studies that we included are realistically powered to identify small, medium, and large scale differences between case and control. When assessing the power of each study at different effect sizes the majority of studies for both adenoma and carcinoma have an 80% power to detect a 30% difference [Figure 7A & B]. No single study that was analyzed had the standard 80% power to detect an effect size difference that was equal to or below 10% [Figure 7A & B]. In order to achieve adequate power for small effect sizes, studies would need to recruit over 1000 individuals for each arm [Figure 7C].

Discussion

Our study identifies clear differences in diversity both at the community level and within individual genera that are present in individuals with CRC versus those without the disease [Figure 1-3]. Although there was a step-wise decrease in diversity from control to adenoma to carcinoma, this did not translate into large effect sizes for the relative risk of either of these two conditions. Even though CRC-associated genera increase the RR of carcinoma, they do not consistently increase the relative risk of adenoma. This information suggests that these specific genera are important in carcinoma genesis but may not be the primary members of the microbial community contributing to the formation of an adenoma. Additionally, our data show that by using the whole community, our models perform better than when only the CRC-associated genera are included. CRC-associated genera are clearly important to carcinoma pathogenesis but accounting for the community in which these microbes exist can drastically increase the ability of models to make predictions. These observations suggest that small localized changes within the community on tissue may be occuring that are important in early disease progression of CRC and that this process may not directly involve CRC-associated genera.

The data presented herein supports the driver-passenger model of the microbial role in CRC, as summarized by Flynn [2], when applied to carcinoma but not necessarily adenoma. The central idea of the model is that a single bacterium initiates an environment in which other non-resident microbes may then be able to colonize, creating a vicious cycle that is conducive for CRC. Both the drastically increased RR of CRC-associated genera versus α-diversity metrics for carcinoma and increasing RR with more CRC-associated generea positivity are highly supportive of this model. However, the initial establishment of the driver within the system appears to be dependent on the current community. This is supported by our finding that when adding the community context to our models in addition to the CRC-associated genera, the model AUC increases [Figure 4 & 5]. Conversely, using the present data, it is less likely that adenoma development fits this model. The changes that occur at this timepoint are small and possibly focal to the adenoma itself. The stepwise decrease in diversity suggests that the adenoma community is not normal but has changed subtly [Figure 1]. Although there appears to be localized changes that do depend on the driver-passenger model, as supported by an increased RR for one, two, and four positive CRC-associated genera in tissue [Figure 3C], there may be other processes at play that ultimately exacerbate the condition from a subtle localized change, to a change in the global community. The poor performance of the Random Forest models for classifying adenoma based only on the microbiota would suggest that this is the case. One potential hypothesis from these observations is that at early stages of the diease, how the host interacts with these subtle changes is what ultimately leads to a thoroughly dysfunctional community that is supportive of CRC genesis.

Within stool, common resident microbes were most consistently present in the top 10 genera or OTUs across study [Figure 6]. Changes in *Bacteroides*, *Ruminococcus*, and *Roseburia* were consistently found to be disciminative across the different studies for both adenoma and carcinoma models [Figure 6]. This data would suggest that whether the non-resident bacterium is *Fusobacteria* or *Peptostreptococcus* is not as important as how these bacteria interact with the changing resident community. These observations would also suggest that initial changes in the resident community, specifically to *Ruminococcaceae*, *Ruminococcus*, and *Roseburia*, carry on from adenoma to carcinoma. Based on these observations, it is possible to hypothesize that the early changes in the community may give rise to initial polyp formation via interactions with the host and not necessarily via interactions with CRC-assciated genera. These changes then create new niches in which any one of the CRC-associated genera could gain a foothold, exacerbating the initial changes in community and facilitating the transition from adenoma to carcinoma via a driver-passenger type mechanism.

The tissue studies did not provide a clearer understanding of how the microbiota may drive the progression of disease severity. For the OTU models both the unmatched and matched [Figure S7E & F] tissue samples had some concordance with the stool data, with resident bacteria being the most prevalent in the top 10 important variables across studies. Unlike in stool, *Fusobacterium* was the only CRC-associated bacteria consistently present in the top 10 of the CRC models [Figure S7B-C & E-F]. The majority of the results seem to be study specific with many top 10 taxa being present only in a single study. One other potentially worrying sign is the presence of *Propionibacterium* within the top 10 for the genera and OTU models and could be a marker of contamination. The low biomass of these samples coupled with potential contamination might be a possible reason why the tissue results seem to be more sporadic then the stool results.

Although there are still questions that need to be answered regarding the microbiota and carcinoma, a clearer framework for their relationship is beginning to develop as to how this occurs. From our observations many changes in carcinoma could easily result in effect sizes that are 30% or more between the case and control and most studies analyzed have sufficient power to detect these types of changes [Figure 7]. Conversely, the associations between the microbiota and adenoma is less clear and part of the reason this may be is because many studies are not powered effectively to observe the small changes reported here. None of the studies analyzed were properly powered to detect a 10% or lower change between case and controls. This small effect size range may well be the scope in which differences consistently occur in adenoma due to the subtle changes in community that occur between control and adenoma. Future studies investigating adenoma and the microbiota need to take these factors into consideration if we are to work out how the microbiota contributes to adenoma formation.

Conclusion

By aggregating together a large collection of studies from both feces and tissue, we are able to provide evidence in support of the importance of the bacterial community in both adenoma and carcinoma. Overall, our results support a framework by which early localized community changes give rise to polyp formation. With the host as a potential catalyst, new niches arise by which non-resident CRC-associated microbes can then gain a foothold and create an environment that allows more of these microbes to colonize. This exacerbates the existing community changes and creates a vicious cycle conducive of carcinoma formation. Our observations also highlight the importance of power and sample number considerations when undertaking investigations into the microbiota and adenoma due to the subtle changes in the community. Although there are power limitations associated with adenoma, this report highlights the strong influence the microbiota has on CRC development.

Methods

Obtaining Data Sets: Studies used for this meta-analysis were identified through the review articles written by Keku, *et al.* and Vogtmann, *et al.* [29,30]. 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 16S rRNA gene sequencing analysis were included. Studies using either 454 or Illumina sequencing technology were included. Only data sets that had sequences available for analysis were included. Some studies did not have publically 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.* [26], Baxter, *et al.* [16], Brim, *et al.* [22], Burns, *et al.* [27], Chen, *et al.* [19], Dejea, *et al.* [24], Flemer, *et al.* [12], Geng, *et al.* [28], Hale, *et al.* [18], Kostic, *et al.* [15], Lu, *et al.* [21], Sanapareddy, *et al.* [25], Wang, *et al.* [20], Weir, *et al.* [23], and Zeller, *et al.* [17]. The Zackular [31] study was not included becasue the 90 individuals analyzed within the study are contained within the larger Baxter study [16]. Additionally, after sequence processing all the case samples for the Kostic study only had 100 sequences remaining and was not used. This left a total of 14 studies for which analysis could be completed.

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

Sequence Processing: For the majority of studies, raw sequences were downloaded from the Sequence Read Archive (SRA) (ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/) and metadata was obtained from the by searching the respective accession number of the study 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 for one study [26] and for four studies was obtained directly from the authors [12,18,23,25]. Each study was processed using the mothur (v1.39.3) software program [32]. Where possible, quality filtering utilized the default methods used in mothur for either 454 or Illumina based sequencing. If it was not possible to use these defaults, the stated quality cut-offs were used instead. Chimeras were identifed and removed using VSEARCH [33] before *de novo* OTU clustering at 97% similarity using the OptiClust algorithm [34] was utilized.

Statistical Analysis: All statistical analysis after sequence processing utilized the R (v3.4.2) software package [35]. For the α-diversity analysis, values were power transformed using the rcompanion (v1.10.1) package [36] and then Z-score normalized using the car (v2.1.5) package [37]. Testing for α-diversity differences utilized linear mixed-effect models created using the lme4 (v1.1.14) package [38] to correct for study and variable region effects in feces and study, variable region, and individual effects in tissue. Relative risk was analyzed using both the epiR (v0.9.87) and metafor (v2.0.0) packages [39,40] by assessing how many with and without disease were above and below the overall median value within the specific study. Relative risk significance testing utilized the chi-squred test. β-diversity differences utilized a Bray-Curtis distance matrix and PERMANOVA executed with the vegan (v2.4.4) package [41]. Random Forest models were built using both the caret (v6.0.77) and randomForest (v4.6.12) packages [42,43]. Differences between the obtained AUC versus a random model AUC was assessed using T-tests. Power analysis and estimations were made using the pwr (v1.2.1) and statmod (v1.4.30) packages [44,45]. All figures were created using both ggplot2 (v2.2.1) and gridExtra (v2.3) packages [46,47].

Study Analysis Overview: α-diversity was first assessed for differences between controls, adenoma, and carcinoma. We analyzed the data using linear mixed-effect models and relative risk. β-diversity was then assessed for each inidividual study. Next, four specific CRC-associated genera (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, and *Porphyromonas*) were assessed for differences in relative risk. We then built Random Forest models based on all genera or the select CRC-associated genera. The models were trained on one study then tested on the remaining studies for every study. The data was split between feces and tissue samples. Within the tissue groups the data was further divided between samples from the same individual (matched) and those from different individuals (unmatched) tissue samples. Where applicable for each study, predictions for adenoma and carcinoma were tested. This same approach was then applied at the OTU level with the exception that instead of testing on the other studies, a 10-fold cross validation was utilized and 100 different models were created based on random 80/20 splitting of the data to generate a range of expected AUCs. For OTU based models, the CRC associated genera included all OTUs that had a taxonomic classification to *Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, or *Porphyromonas*. The power of each study was assessed for an effect size ranging from 1% to 30%. An estimated sample n for these effect sizes was also generated based on 80% power. For comparisons in which normal versus adenoma were made the carcinoma samples were excluded from each respective study. Similarily, for comparisons in which normal versus carcinoma were made the adenoma samples were excluded from each respective study.

Obtaining CRC-Associated Genera: For the CRC-associated genera analysis of the RR, the total average counts were collected for each respective OTU that had a genus level taxonomic classification to *Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, and *Porphyromonas* for 100 different subsamplings. The OTU based Random Forest Models that used CRC-associated genera used a similar approach except that the OTUs were not aggregated together by genus by kept as separate OTUs. So, OTU Random Forest models using the full community included all OTUs while those using CRC-associated genera included only those OTUs that had a genus level taxonomic classification to *Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, and *Porphyromonas*.

Matched versus Unmatched Tissue Samples: In general, tissue samples that had control and lesion samples that did not belong to the same indivdiual 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. Studies that had unmatched data were Burns, Flemer, Chen, and Sanapareddy. For some studies samples became unmatched due to one of the corresponding matched samples not making it through sequence processing. For the linear mixed-effect models samples from the same individual were taken into account. For all other analysis matched and unmatched samples were analyzed separately using the statistical approaches mentioned in the Statistical Analysis section.

Assessing Important Random Forest Model Variables: The genus based models collected the top 10 most important variables from each training set and assessed how many times that genera showed up in the top 10 across each study. The OTU based models recorded the medians for each OTU across 100 different 80/20 splits of the data for each study. The lowest classification for each OTU was obtained using the RDP database and the number of times the specific classification occured in the top 10 across studies was recorded. For the adenoma tissue genus and OTU models there was only one matched and unmatched study and these results were grouped together for the counting of the top 10.

Reproducible Methods: The code and analysis can be found here https://github.com/SchlossLab/Sze_CRCMetaAnalysis_Microbiome_2017. Unless mentioned otherwise, the accession number for the raw sequences for 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 Candian Institute of Health Research fellowship and a University of Michigan Postdoctoral Translational Scholar Program grant.

Authors' contributions

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

15. 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.

16. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine.

2016;8:37.

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

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

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

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

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

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

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. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL, Rossetti BJ, et al. Microbiota organization is a distinct feature of proximal colorectal cancers. Proceedings of

the National Academy of Sciences of the United States of America. 2014;111:18321–6.

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

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

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

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

29. 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

30. 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

31. 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.

32. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl.Environ.Microbiol. [Internet].

2009 [cited 12AD Jan 1];75:7537–41. Available from: http://aem.asm.org/cgi/content/abstract/75/23/7537

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

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

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

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

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

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

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

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

i03/

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

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

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

44. Champely S. Pwr: Basic functions for power analysis [Internet]. 2017. Available from: https://CRAN.R-project.org/package=pwr

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

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

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

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

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

Table 2: Studies with Tissue Samples Included in the Analysis

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

Figure 1: α-Diversity Differences between Control, Adenoma, and Carcinoma Across Sampling Site. A) α-diversity metric differences by group in stool samples. B) α-diversity metric differences by group in unmatched tissue samples. C) α-diversity metric differences by group in matched tissue samples. The dashed line represents a Z-score of 0 or no difference from the median.

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

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

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

Figure 5: OTU Random Forest Model of Tissue Across Studies. A) Adenoma random forest model between the full community and CRC-associated genera OTUs only. B) Carcinoma random forest model between the full community and CRC-associated genera OTUs only. The dotted line represents an AUC of 0.5 and the lines represent the range in which the AUC for the 100 different 80/20 runs fell between. The solid red line represents

the median AUC of all the studies for either the full community or CRC-associated genera OTUS only model.

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

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

Figure S1: Relative Risk for Adenoma or Carcinoma based on α-Diversity Metrics in Tissue. A) α-metric relative risk for adenoma. B) α-metric relative risk for carcinoma. Colors represent the different variable regions used within the respective study.

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

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

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

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

Figure S6: Random Forest Prediction Success of Adenoma Using Genera for each

Tissue Study. The red lines represent the median AUC using that specific study as the training set.

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