The Microbiota and Individual Community Members in Colorectal Cancer: Is There a Common Theme?

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.

Results.

Conclusions.

Keywords

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

Background

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

Using both fecal and tissue samples totalling over 2100 total individuals across 14 studies within our data analysis we exapnd 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. We first assessed the diversity of controls, adenoma, and carcinoma individuals and tested whether they change and if it results in an increased relative risk of adenoma or carcinoma. Next, we assessed how this relative risk compared to CRC associated genera for both adenoma and carcinoma. Third, using Random Forest models we assessed whether the community context can increase the classification model area under the curve (AUC). Finally, we examine whether the studies that were used were adequately powered and if not what effect size they were powered for.

Our analysis found a continuous decreae in Shannon diversity from control to adenoma to carcinoma and a significantly increased relative risk for carcinoma with lower diversity. Using the CRC associated genera only this relative risk was higher than Shannon diversity. However, adding the community context in which these CRC associated genera are present increases prediction models AUC. Although we analyze a data set with a large number of

total individuals each individual study was underpowered for effect size differences of 10% or below between the case and control.

Results

Fecal Diversity is Lower in Those with Carcinoma and Increases Relative Risk: Using power transformed and Z-score normalized alpha diversity metrics both evenness and the Shannon diversity metrics in feces are lower in those with carcinoma then in controls but not for tissue samples [Figure 1]. Using linear mixed-effects 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). This effect was not observed in tissue when additionally controlling for whether the sample came from the same individual (P-value > 0.05). For fecal samples a decrease in Shannon diversity and evenness resulted in a significantly increased relative risk for carcinoma (P-value = 0.01 and P-value = 0.0011, respectively) [Figure 2]. Although these values were significant the effect size was relatively small for both metrics (Shannon RR = 1.31 and evenness RR = 1.34) [Figure 2]. There was no increased relative risk for these metrics for adenoma or for tissue in general [Figure S1-3].

Using the Bray-Curtis distance metric, the fecal microbiota did not have a different community diversity between adenoma and control but did for carcinoma across studies [Table S1 & S2]. The majority of unmatched tissue samples had a significant difference for both adenoma and carcinoma versus controls [Table S3 & S4]. All matched tissue samples accross studies had no difference between any of the compared groups [Table S3 & S4].

Genera Previously Associated with Carcinoma Increases Relative Risk More than Alpha Diversity: Both fecal and tissue samples had a significantly increased RR for carcinoma but not for adenoma [Figure 3] which was greater than either evenness or Shannon diversity [Figure 2 & 3]. The relative risk did not increase when considering the total abundance or increasing number of carcinoma associated genera [Figure 3]. The RR effect size was greater for stool (RR range = 1.78 - 2.64) then for tissue (RR range = 1.33 -

1.53). This decrease may be explained by the fact that tissue samples include matched samples.

Using the Whole Community Increases Model AUC over CRC Associated Genera: For both fecal and tissue samples (matched and unmatched) there was a decrease in AUC when only OTUs from the CRC associated genera are used [Figure 4 & 5]. This decrease is observed in both adenoma and carcinoma groups [Figure 4 & 5]. The genus models generally had similar trends as observed for the OTU based models with the full genera models performing better then the CRC associated genera models [Figure S4-S5]. Both genus models perform similarily in their ability to be able to predict lesion (adenoma or carcinoma) with carcinoma having a higher AUC then adenoma [Figure S6-S8]. Matched tissue samples for those with carcinoma had an AUC that was more similar to the adenoma models [Figure S6A, S7B, & S8] then carcinoma models [Figure S6B & S7A].

Majority of Studies are Underpowered for Detecting Small Effect Size Differences: 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 6A & B]. No single study that was analyzed had the standard 80% power to detect a difference that was eqaul to or below 10% [Figure 6A & B]. In order to achieve adequate power for small effect sizes it would be necessary to recruit over 1000 individuals for each arm of the study [Figure 6C].

Discussion

Our study identifies clear diversity changes both at the community level and within individual genera that are present in indivdiuals with carcinoma versus those without the disease. 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. These clear changes were not easily recapitulated in those with adenoma. Even though CRC associated genera increase the relative risk of carcinoma they do not increase the relative risk of adenoma. This information suggests that these specific genera may not be the primary members of the microbial community that contributes to the formation of an adenoma but is for a carcinoma. Additionally, our data shows that by using the whole community our models perform better then when they only use the CRC associated genera. CRC associated genera are clearly important to carcinoma but the context or community in which these microbes are a part of can drastically increase the ability of models to make predictions. This data supports the concept that small localized changes within the community may be occuring that are important in the disease progression of colorectal cancer and that they may not directly involve CRC associated genera.

The driver-passenger model of the microbial role in CRC, as summarized by Flynn [1], can be supported with this data for carcinoma but not necessarily for adenoma. The drasitically increased relative risk of disease when considering the CRC associated genera is highly supportive of this type of process. In a driver-passenger scenario it is possible that simply having the driver present or only identifying the passenger is a good enough proxy that the event is occuring. This would account for the observation that there is no synergistic increase in relative risk when accounting for either the total number or increasing abundance of these genera. The initial establishment of the driver within the system is also dependent on the community that is present and this is supported by the

observation that when adding the community context to our models along with the CRC associated genera the model AUC increases.

Our carcinoma observations fit the driver-passenger model and support this concept within the framework of the transition from adenoma to carcinoma. In contrast, the adenoma observations do not fit well with this model and suggests that the transition from control to adenoma do not fit this framework. The stepwise decrease in diversity suggests that the adenoma community is not normal but this change is subtle. Although there may be localized changes that do depend on the driver-passenger model, our observations show that they do not involve the CRC associated genera. It is possible to hypothesize that at early stages of the diease, how the host interacts to these subtle changes could be the catlyst that causes adenoma formation. Subsequent transition to carcinoma could then fit into the proposed driver-passenger model framework.

Although there are still questions that need to be answered for the microbiota and carcinoma a clearer framework is beginning to develop as to how this occurs. The role of the microbiota in adenoma is still not clear and part of the reason may be because many studies are not powered effectely to observe the small changes reported here. It is realistic to suspect that many changes in carcinoma could easily result in effect sizes that are 30% or more between the case and control. Most of the studies analyzed have sufficient power to detect these changes. In contrast, our data suggests that the adenoma effect size is relatively small. None of the studies analyzed were properly powered to detect a 10% or lower change between case and controls and this may well be the range in which differences occur in adenoma. Future studies investigating adenoma and the microbiota need to take these factors into consideration if we are to work out the role of the microbiota in adenoma genesis.

Conclusion

By aggregating together a large collection of studies from both feces and tissue we are able to provide information in support of the driver-passenger model in the context of carcinoma. However, within the context of adenoma it is less clear that this relationship exists. These observations highlight the importance of power and sample number considerations when considering investigations into the microbiota and adenoma due to the subtle changes in community. This study helps to identify the problems that have been solved and the challenges that lie ahead in the invesitgation of colorectal cancer and the microbiota.

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. [2,3]. All studies were included that used tissue or feces as their sample source for 16S rRNA gene sequencing analysis. Studies using either 454 or Illumina sequencing technology were included. Only data sets that had the raw sequences available for analysis were included. Some studies did not have publically available raw sequences or did not have meta data in which the authors were able to share. After this filtering step the following studies remained: Ahn [4], Baxter [5], Brim [6], Burns [7], Chen [8], Dejea [9], Flemer [10], Geng [11], Hale [12], Kostic [13], Lu [14], Sanapareddy [15], Wang [16], Weir [17], and Zeller [18]. The Zackular [19] study was not included becasue the 90 individuals analyzed within the study are contained within the larger Baxter study. The Kostic study was not used since after sequence processing all the case samples did not have more than 100 sequences remaining. This left a total of 13 studies in which complete analysis could be completed.

Data Set Breakdown: In total there were 7 studies with only fecal samples (Ahn, Baxter, Brim, Hale, Wang, Weir, and Zeller), 5 studies with only tissue samples (Burns, Dejea, Geng, Lu, Sanapareddy), and 2 studies with both fecal and tissue samples (Chen and Flemer). The total number of individuals initially run through the sequence processing for the fecal samples was 1899 and for the tissue samples was 462.

Sequence Processing: For the majority of studies raw sequences were downloaded from the SRA (ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/) and metadata was obtained from the following website: http://www.ncbi.nlm.nih.gov/Traces/study/ by searching the respective accession number of the study. Of the studies that did not have sequences and meta data on the SRA one study had the data stored on DBGap [4] and four studies the data was obtained directly from the authors [10,12,15,17]. Each

study was processed using the mothur (v1.39.3) software program [20]. 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 author stated quality cut-offs were used instead. Chimeras were identifed and removed using the VSEARCH [21] program and *de novo* OTU clustering at 97% similarity using the OptiClust algorithm [22] was utilized.

Statistical Analysis: All statistical analysis after sequence processing utilized the R software package (v3.4.2). For the alpha diversity analysis values were power transformed using the rcompanion (v1.10.1) package and then Z-score normalized using the car (v2.1.5) package. Testing for alpha diversity differences utilized linear mixed-effect models created using the lme4 (v1.1.14) package to correct for both study and variable region effect in the diversity measures when analyzing colorectal cancer groups. Relative Risk was analyzed using both the epiR (v0.9.87) and metafor (v2.0.0) packages. Relative risk significance testing utilized the chi-squred test. Beta-diversity differences utilized a Bray-Curtis distance matrix and PERMANOVA executed with the vegan (v2.4.4) package. Random Forest models were built using both the caret (v6.0.77) and randomForest (v4.6.12) packages. Random Forest testing of the obtained AUC versus a random model AUC utilized T-tests. Power analysis and estimations were made using the pwr (v1.2.1) and statmod (v1.4.30) packages. All figures were created using both ggplot2 (v2.2.1) and gridExtra (v2.3) packages.

Study Analysis Overview: Alpha diversity was first assessed for differences between controls and adenoma versus cancer and controls versus adenoma. We analyzed the data using linear mixed-effect models, and relative risk. Beta-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 matched and unmatched tissue samples. Both prediction 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 and effect size ranging from 1% to 30%. An estimated sample n for these effect sizes was also generated based on 80% power.

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, on 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 manuscript used in this meta-analysis.

Consent for publication

Not applicable.

Availability of data and material

A detailed and reporducible 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 studies batch file in the GitHub repo or within the original publication. When sequences were not publicly available contacting the corresponding author for raw sequences needs to be undertaken.

Competing Interests

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

Funding

MAS is supported by a CIHR fellowship and a University of Michigan PTSP fellowship 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 apart of each of the individual studies uitlized. 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. Flynn KJ, Baxter NT, Schloss PD. Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. mSphere. 2016;1.

2. 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

3. 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

4. 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.

5. 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.

6. 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.

7. 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.

8. Chen W, Liu F, Ling Z, Tong X, Xiang C. Human intestinal lumen and mucosa-associated

251 microbiota in patients with colorectal cancer. PloS One. 2012;7:e39743.

252 9. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL, Rossetti BJ, et al.
253 Microbiota organization is a distinct feature of proximal colorectal cancers. Proceedings of
254 the National Academy of Sciences of the United States of America. 2014;111:18321–6.

255 10. Flemer B, Lynch DB, Brown JMR, Jeffery IB, Ryan FJ, Claesson MJ, et al.
256 Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut.
257 2017;66:633–43.

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

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

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

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

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

272 16. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut
273 microbiota between colorectal cancer patients and healthy volunteers. The ISME journal.

2012;6:320–9.

17. 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.

18. 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.

19. 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.

20. 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

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

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

Table 1:

Figure 1:

Figure 2:

Figure 3:

Figure 4:

Figure 5:

Figure 6:

Figure S1:

Figure S2:

Figure S3:

Figure S4:

Figure S5:

Figure S6:

Figure S7:

Figure S8: