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. An increasing body of literature suggests that there is a role for the microbiota in colorectal cancer. Important drivers within this axis have ranged from individual microbes to the whole community. A recent meta-analysis investigated whether consistent biomarkers could be identified across studies. This report expands on this previous research and tests the hypothesis that the bacterial community is an important driver of disease from control to adenoma to carcinoma. To test this hypothesis we examined both feces (total individuals = 1737) and tissue (total samples = 492) across 14 different studies.

Results. Fecal samples had a significant decrease from control to adenoma to carcinoma for both Shannon diversity and evenness (P-value < 0.05) after correcting for study effect and variable region sequenced. Lower Shannon diversity and evenness in fecal samples resulted in a significant increase in relative risk for carcinoma (P-value < 0.05) while only evenness for adenoma (P-value < 0.05) resulted in a slightly increased relative risk. Previously associated colorectal cancer genera (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, or *Porphyromonas*) followed a similar pattern with a significantly increased relative risk by their presence for carcinoma (P-value < 0.05) but not adenoma (P-value > 0.05) with the exception of *Porphyromonas* (P-value < 0.05). Using the whole community versus only CRC associated genera to build a prediction model resulted in a higher classification AUC for both adenoma and carcinoma for fecal and tissue samples. For the studies that were analyzed, most were adequately powered for large effect size differences which may be more amenable for carcinoma than for adenoma microbiota research.

Conclusions. This data provides support for the importance of the bacteral community to both adenoma and carcinoma genesis. The evidence that has been collated within

this study is much stronger for carcinoma and 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 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 [11]. 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 (total individuals = 1737) and tissue samples (total samples = 492) totalling over 2229 total samples across 14 studies [12–25] 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 across disease (control to adenoma to carcinoma) and if it results in an increased relative risk for adenoma or carcinoma. Next, we assessed how typical CRC associated genera (*Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, or *Porphyromonas*) affect the relative risk of adenoma or carcinoma. Third, using Random Forest models we analyzed whether the full community or only the CRC associated genera resulted in better model classification area under the curve (AUC). Finally, we examined at what effect size and sample n the studies that were used were adequately powered for.

Our analysis found a continuous decrease in Shannon diversity from control to adenoma to carcinoma and a significantly increased relative risk for carcinoma with this lower diversity. Using only the CRC associated genera this relative risk was higher than what was observed for Shannon diversity. However, by using the full community instead of only

the CRC associated genera the AUC of the classification models increased. 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 group.

Results

Diversity is Lower in Those with Carcinoma and Increases Relative Risk: Using power transformed and Z-score normalized α-diversity metrics, both evenness and Shannon diversity in feces and 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). 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]. Only a decrease in evenness had an increased relative risk 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).

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 between adenoma and controls [Table S1 & S2]. Many studies with unmatched tissue samples had a significant difference for both adenoma and carcinoma versus controls [Table S3 & S4] while studies with matched tissue samples had no differences [Table S3 & S4].

Genera Previously Associated with Carcinoma Predominately Increases Relative Risk for Carcinoma but not Adenoma: The majority of CRC associated genera for both feces and tissue had a significantly increased relative risk for carcinoma but not for adenoma [Figure 3]. In fecal samples the relative risk due to CRC associated genera was greater than either evenness or Shannon diversity [Figure 2 & 3]. Additionally, the

relative risk of carcinoma continuously increased as individuals tested positive for more CRC associated genera [Figure 3B & 3D]. The relative risk effect size was greater for stool (RR range = 1.61 - 2.74) then for tissue (RR range = 1.21 - 1.81). This decrease may be explained by the fact that the tissue analysis include matched samples.

Two measures in stool for adenoma were significant when investigating these CRC associated genera. The first was *Porphyromonas* (P-value = 0.023) and the second was being positive for three CRC associated genera (P-value = 0.022) [Figure 3A]. For tissue three measures for adenoma were significant; being positive for one CRC associated genera (P-value = 0.032), being positive for two CRC associated genera (P-value = 0.008), and being positive for four CRC associated genera (P-value = 0.039) [Figure 3C].

Using the Whole Community Increases Model AUC over CRC Associated Genera: For both fecal and tissue samples (matched and unmatched) the AUC decreases 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 S2-S3]. 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 S4-S5]. Matched tissue samples for those with carcinoma had an AUC that was more similar to the adenoma models [Figure S4A, S5B, & S6] then carcinoma models [Figure S4B & S5A].

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 an effect size 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 individuals 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. Even though CRC associated genera increases the relative risk of carcinoma it does 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 that contribute to the formation of an adenoma. 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 [2], 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, especially in the context of increasing relative risk with more CRC associated genera positivity. It is also possible that in a driver-passenger scenario 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 constant additive effect on relative risk for increasing positivity. Additionally, 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, with the present data we can only suggest that the adenoma observations might fit with this model but the changes that occur at this timepoint are small and possibly focal to the adenoma. 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, supported by an increased relative risk for one, two, and four positive CRC associated genera in tissue [Figure 3C], there may be other processes involved that ultimately exacerbates the condition from a subtle localized change to a global community one. The poor performance of the Random Forest models for classifying adenoma based only on the microbiota would suggest that this is the case. 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 ultimately leads to this larger global dysfunctional community.

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 this may be is because many studies are not powered effectively 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 type of 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 consistently 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 formation.

Conclusion

By aggregating together a large collection of studies from both feces and tissue we are able to provide information in support of the importance of the bacterial community in both adenoma and carcinoma. We are also able to provide support for 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 the community.

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.* [26,27]. 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.* [21], Baxter, *et al.* [24], Brim, *et al.* [17], Burns, *et al.* [22], Chen, *et al.* [14], Dejea, *et al.* [19], Flemer, *et al.* [13], Geng, *et al.* [25], Hale, *et al.* [12], Kostic, *et al.* [28], Lu, *et al.* [16], Sanapareddy, *et al.* [20], Wang, *et al.* [15], Weir, *et al.* [18], and Zeller, *et al.* [23]. The Zackular [29] study was not included becasue the 90 individuals analyzed within the study are contained within the larger Baxter study [24]. 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 [21] and for four studies was obtained directly from the authors [12,13,18,20]. Each study was processed using the mothur (v1.39.3) software program [30]. 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 [31] before *de novo* OTU clustering at 97% similarity using the OptiClust algorithm [32] was utilized.

Statistical Analysis: All statistical analysis after sequence processing utilized the R software package (v3.4.2). For the α-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 α-diversity differences utilized linear mixed-effect models created using the lme4 (v1.1.14) package 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. 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. Random Forest models were built using both the caret (v6.0.77) and randomForest (v4.6.12) packages. 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. All figures were created using both ggplot2 (v2.2.1) and gridExtra (v2.3) packages.

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 matched and 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.

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 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. 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. 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.

13. 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.

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

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

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. Geng J, Fan H, Tang X, Zhai H, Zhang Z. Diversified pattern of the human colorectal

cancer microbiome. Gut Pathogens. 2013;5:2.

26. 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

27. 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

28. 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.

29. 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.

30. 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

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

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

Table 1: Studies with Stool Samples Included in the Analysis

Study	Data Stored	16S Region	Controls	Adenoma	Carcinoma
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	Controls	Adenoma	Carcinoma
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: Colorectal Cancer Associated Genera Relative Risk for Adenoma and Carcinoma in Stool and Tissue. A) Adenoma relative risk in stool. B) Carinoma 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.

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.

Figure 6: 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

386 with Carcinoma individuals. C) The estimated sample number needed for each arm of
387 each study to detect aneffect size of 1-30%. The dotted red lines in A) and B) represent
388 the generally used 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 CRC associated genera only 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.

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.

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