We would like to thank the editor and reviewers for taking the time to provide their thoughtful and valuable feedback. These constructive comments greatly improved our manuscript.

There were two primary concerns in the reviewer feedback. The **first** was around the wording and descriptions, which we addressed in our overall tone, clarity in the methods sections, and interpretation/discussion of the results. The **second** primary concern was related to our data in controlling for contamination (negative controls, validation of phage sequence sources, etc). We provided further explanation and data around the negative (background) controls that we used, as well as more clearly validated that we are measuring a real signal. This included the addition of filtering steps and re-running all of our analyses with more conservatively filtered data. Overall we present a manuscript with modified text and figures.

Below we provide a point-by-point response to the reviewer comments. Our changes in the manuscript are marked in red and blue text.

Reviewer #1 (Comments for the Author):

Article Title: The Colorectal Cancer Virome

This article describes work describing associations between bacterial and viral communities with colorectal cancer. The authors used a mixture of techniques (16S rRNA amplicon profiling, whole-shotgun metagenomics and purified virus metagenomic sequencing) to characterize the bacterial and viral components of the enteric microbiome of colorectal patients and controls. Detailed analysis, relying heavily on random-forest decision trees were used to identify differences in healthy and cancer associated viromes. The results indicate that there are indeed differences in the viromes of colorectal cancer patients that distinguish them from 'healthy' controls. The differences are primarily in the abundances of temperate bacteriophages suggesting that phage may be contributing to microbiome remodeling via predator-prey interactions of phage and bacteria.

Overall the authors have generated a very interesting and likely very important data set for understanding the complex microbial interactions associated with, and potentially driving, colorectal cancer development. However, I believe there are some critical flaws to the study as detailed below. In summary, these flaws are 1) inadequate conceptual delineation between direct and indirect effects of the virome. This is largely a prose problem, is addressed in specific portions but as written is conceptually confusing, 2) over-stating and under-valuing the lack of considering the full virome. The lack of RNA and enveloped virus analysis should dampen some of the conclusions made, 3) inadequate pursuit on the classification of OVUs with importance in many of the models and 4) overstated model. Points 1, 2 and 4 can be handled with adjustments to the tone and additional consideration of alternative interpretations. Point 3 requires additional analysis wherein

greater efforts are made (some suggestions below) to identify the nature of the contigs in the discriminatory OVUs and to confim that they are not contamination from bacterial (or other) chromosomes.

Thank you very much for the feedback and well articulated points of revision. We agree that revisions around these points will greatly improve the manuscript.

- Much of the premise is about how viruses can cause cancer, but there is no evidence from the literature (that I am aware of) or evidence provided in this manuscript to indicate that bacteriophage can directly induce cancer. Instead, the premise is that bacteriophage are modulating bacteria which may result in changes associated with or potentially causing cancerous state. Statements like this: Due to their mutagenic abilities and propensity for functional manipulation, human viruses are strongly associated with, and in many cases cause, cancer" confuse the interpretation of the results and the manuscript would be more clear if the concept of viruses directly causing cancer were isolated and conceptually distinct from the phage-bacteria components. This is particularly confusing as the study only analyzed the DNA virome and could be missing a huge number of viruses that may be directly involved in cancer formation or acceleration.

  We agree that this needs to be clearer. When we used the term "human viruses" we were referring to viruses that only infect humans, and not bacteriophages. This is confusing because the human virome consists of both of these types of viruses, and ultimately this paper focuses on the bacteriophage component because that is the direction the data took us. We re-worded this section of the text to better differentiate those viruses that we would expect to indirectly and directly impact cancer development. **LINES 19, 33, 41-54, 225-227**

- What defines a "healthy human colon"? It is normal practice that "healthy control" samples obtained for studies like this are actually from patients who enter into the clinic for a different disorder. It is unclear how colons would be taken from 'healthy' individuals as this is not normal. The authors need to describe the criteria defining 'healthy'. The section "Study Design and Patient Sampling". Should include detailed information about controls.

  Our first point of clarification here is that this study is based on stool samples and not colon biopsies or similar sample types (referring to the comment "It is unclear how colons would be taken from 'healthy' individuals as this is not normal") and we added this point more frequently in the text. Second, we defined healthy as "had not had surgery, had not had chemotherapy or radiation, and were free of known co-morbidities including HIV, chronic viral hepatitis, HNPCC, FAP, and inflammatory bowel disease" as well as confirmation by a colonoscopy, as is outlined

in the methods section of the manuscript. Patients were recruited specifically for this study and were not recruited upon presentation with co-morbidities. In order to minimize the table count and redundancy of tables, we referenced the reader to the original paper that used this set of stool samples. We are happy to move those previously published data to tables within our manuscript if the editor and reviewers feel it is necessary. **LINES 86, 301-310**

- The statement "As expected, these controls yielded few sequences and were almost entirely removed while rarefying the datasets to a common number of sequences (Figure S1 B)." is unclear as the barplots in S1B show the emergence of more detectable CLP sequence yield in control (blue) samples. The figure legend to S1 doesn't mention rarefaction and would benefit from further explanation including what sequence number the data were rarified instead of relying on digging into the methods for this information.

  We were attempting to communicate that we subsampled all of our sequences down to 1 million reads and removed all samples with a sequencing depth below that count. We adjusted the plot to show the rarefaction depth and illustrate the elimination of samples. Our rarefaction process resulted in all but the outlier negative control samples being removed, as well as a small fraction of samples from the other experimental groups. We had originally included a quantification of the dissimilarity of our rarefied samples to support the difference between the negative control samples and the other experimental samples, along with a depth for statistical significance. **FIGURES S01, S05, LINES 105-116**

- The results section on the advantages of viral methods to bacterial (16S and metagenomic) should be rewritten. The results are interested and intriguing, but the premise is flawed. One should never attempt to "determine the advantage" there are only differences and results. The concluding statement that Despite the recent loss of enthusiasm for 16S rRNA gene sequencing in favor of shotgun metagenomic techniques, 16S rRNA gene sequencing is still a superior methodological approach for some important applications." Is an editorial statement, not a result. If the authors wish to pursue this thought, then they should at a minimum move this editorial to the discussion and provide references. As the data stand now, these statements are a generalization about a very specific set of analysis in a very specific set of conditions.

  We felt this was an interesting observation that we think is worth pointing out to the reader. However we modified out text accordingly, so as to be less polarizing. **LINES 138-154**

- The classification of OVU's as unknown leaves the results somewhat wanting. Is there any additional

information in the contigs of the OVU's? Do they have ORF's? Even minimal information would dramatically change the interpretation and value of these results. In addition, the methods indicate that the OVU's were only queried against viral databases. The virus purification protocols are enrichment protocols meaning that residual nucleic acid from non-viral species can and will be sequenced along with the viral nucleic acid. Therefore, it is very possible that contigs in Unknown Clusters could belong to bacteria, host, fungi, food. However, this information is hidden as the contigs were not queried against a more comprehensive database. This can be computationally expensive, but in particular for the contigs/clusters from Figure 2 C, D and E is absolutely necessary for accurate interpretation of the results. What if cluster 115 is really just contaminating fusobacterium chromosomal nucleic acid?

*We additionally characterized the OVUs by looking at the potential host ranges of the bacteriophages and how this signal generally correlated with importance to the random forest model (**Figure 5**). We also initially ran a loose blast between our virus OGUs and the virus database to identify whether the viruses were bacteriophages, and if so, what kinds. When performing this analysis we found that ~80% of the OVUs were similar to known viruses, and most of those were bacteriophages. We then performed a stricter nucleotide alignment to assign the taxonomic IDs to the OVUs. This criteria was outlined in the manuscript. The larger concern around contamination is noted in the next point, and is addressed in the next point of revision.*

- Similar to the above statement, if contaminating bacterial chromosomes carried through the virion prep protocol how certain are the authors that the podoviridae in S1D (Adenoma vs. Cancer) is not just integrated phage in the bacterial chromosome? Is there any additional information to support it's taxonomy? Any commentary on whether it is related to the porphyromanas in the companion panel G?

There are a couple pieces of evidence to support the lack of contamination in our samples and the strength of the viral signal. First, we performed an additional filtering step, as has been done in previous studies. We aligned the OGUs to both a viral and bacterial reference genome database. We then removed all OGUs which had loose nucleotide similarity to bacteria and no similarity to known viruses. Second, we did not see a predictive signal when we performed the same analysis on the whole metagenomic sample set. We would predict a signal that was bacterial and not viral would show up in both datasets, to even a detectable level, but we did not observe such a signal. Finally, we also looked for correlations between the bacterial and phage sample sets to confirm that there is minimal correlation between the taxonomic abundances which, if observed, would also have suggested the viral signal was actually bacterial. Together this evidence provides support that

the signal was indeed viral. **LINES 392 - 400**

- The lack of correlation between bacteriophage and bacteria relative abundances is not definitive proof of the absence of a predator-prey relationship. Bacteriophage-bacteria communities are complex. A bacteria could become more abundant due to the opening of a niche by bacteriophage predating a bacteria at the same time as being predated by another bacteriophage. The results in the manuscript are important, but are overinterperted and do not make room for alternative interpretations.

  This is a very important point. We did not intend to communicate that the lack of correlation between bacteriophage and bacteria relative abundances is proof of the absence of a predator-prey relationship. Our correlation analysis was performed to provide evidence that the virome was not strictly a reflection of the bacterial community. This provides further interest in the complex community interactions between bacteria and their phages. We appreciate this valuable feedback and reworded our text to make this point clearer, limiting over-interpretation, and allowing room for alternative interpretations. **LINES 202 - 223, 271 - 285**

- The discussion is over-editorialized. Comments such as "more sophisticated" machine learning approaches are not valuable to the community. Alpha and beta diversity measures are also sophisticated and complex in their own ways, and trying to push one over the other does not serve the important findings of this paper well and only serve as fodder for research groups to "pick sides" on which techniques they prefer given their systems.

  We agree that alpha and beta diversity metrics, as well as other approaches for understanding communities, are also extremely valuable and sophisticated. We removed the statement of sophistication from the text. **LINE 238**.

- The lack of analysis of RNA and enveloped viruses should be included in statements such as "The colorectal cancer virome was composed primarily of bacteriophages.". It could very well be that an RNA virus is the most abundant virus in the colorectal cancer virome. This should be made more clear in the discussion.

  This is an important point that has served as a valuable technical caveat for many virome studies over the years. Ours is an initial step in a long road to a highly detailed understanding of the colorectal cancer virome, and we want all of the caveats to be correctly communicated. We made this clearer in our text. **LINES 280 - 285**

- The model is too over-reaching (Figure 5). No evidence is provided in this manuscript for 1) colonization

of diver bacteria, 2) passenger bacteria adherence, 3) biofilm formation, 4) phage-mediated dispersal, 5) linkages of bacteria or phage to epithelial cell transformation, 6) tight-junction disruptions, 7) bacterial infiltration, 8) oncogenic synergy with tumor cells, 9) bacteria "thriving" off secreted peptides, 10) release of carcinogenic reactive oxygen species and polyspermines. These are all great concepts and likely some of them are correct, but this is all just wild speculation at this point with little to no primary data (or literature citations) for many of the points. The authors have assembled an excellent data set with interesting associations between viral and bacterial populations in the context of colon cancer and do not need to wildly speculate on data poor models.

The majority of the model we described was originally detailed in a recent review article by Flynn et al, which provides up-to-date evidence on how bacteria may be influencing colorectal cancer, complete with extensive citations to existing literature. In our manuscript, we wanted to end our manuscript by presenting a hypotheses of ways we think bacteriophages could be playing an additional role in that existing model. It is true that our points are only hypotheses and are meant to represent potential insight into the future directions of our research program and were not meant to represent biological truth as we know it now. We think this is valuable as a "future directions" piece so we opted to keep this in the text. However, we want this model to be properly communicated as a collection of hypotheses for future work, so we revised the text to make this clearer. **LINES 255 - 279, FIGURE 6**

Reviewer #2 (Comments for the Author):

The manuscript by Hannigan et al addresses a topic of great interest - the intestinal virome, and its potential link to human disease. There is existing data on associations between the virome and some conditions such as inflammatory bowel diseases, but this would be the first study of the virome in colon cancer. The central finding of this study was that relative abundances of viral taxa could be used to create random forests classifiers differentiating healthy colon tissue from colon cancer tissue with similar accuracy as microbiome data from 16S rRNA sequencing. This novel finding suggests that the role of phages in colon cancer warrants further investigation. The authors provide detailed description of their methods and access to the raw sequence data and code, which are critical for promoting research efforts in this field.

We appreciate the encouraging words and the support in promoting open access research and reproducible science. This is an important component to our work, and we feel an important component to all science, and we appreciate the feedback.

The main weaknesses of this manuscript are:

1) It overintreprets the classifier findings. Viruses were less able to differentiate adenomas from health or cancer than were bacteria and did not provide any additive predictive power when combined with microbiota to create classifiers. This suggests that the viruses act through the bacteria, but it is not clear that viruses have a causative role through changing the gut microbiome as proposed by the authors, as opposed for instance to being passengers that shift in response to changes in their host microbiota (since most of the viruses were lysogenic). The model in Fig. 5b should be expanded to show alternative explanations in which the phages are markers of changes in bacterial communities associated with colorectal malignancy without being the driver of these changes. The Discussion should also be revised to further address alternative models to explain the phage association with colon cancer (currently it includes just a single sentence on the possibility of direct host recognition of phages).

These are all very important points and we agree that they can be made clearer in our manuscript. We revised the text to make these points clearer, and more accurately describe the scope of our findings and interpretations. **LINES 271 - 285, FIGURE 6**

2) It includes a relatively small number of samples (30 per category) without a validation set. The small sample size may have limited classifier accuracy. This is evident in the large variability in model AUC seen in Fig. S7 for the viral and bacterial classifiers. While this is fine given the lack of an existing literature on this topic, it should be mentioned in the Discussion.

The point around small sample sizes was already observed in the 16S rRNA gene sequencing component of this study. This smaller dataset that we used was the first of a larger study cohort to be analyzed by 16S rRNA gene sequencing methods, and this publication will open new doors for further sampling and access to other sample sets within the cohort. To the point of having a validation dataset, while the samples are all part of one cohort, we used a nested cross validation approach that sets aside a subset of the data to use as the validation set after the model is trained. We mention the scope of sample size in the manuscript. **LINES 271 - 285**

3) There is little characterization of the specific bacteriophage features that differentiate colon cancer from healthy tissue. There was no difference by alpha and beta diversity, and differential abundance testing doesn't seem to have been performed. Random forests classifiers highlighted features that were important for classifier accuracy, but the actual abundances were only shown for 6 features (Fig. S8). It wasn't clear how many were statistically significantly different between healthy and cancer, and little is known about each. It would be helpful if the predicted bacterial hosts for these 6 OVUs (based upon the methods used by the authors to construct their phage-bacteria interaction network) were described

to convey more information about the bacteriophages that may be relevant to colon cancer.

We appreciate this feedback, and there are a few points to mention here. We did provide information about the wide range of hosts predicted for the bacteriophages (correlation between importance and host network centrality, Figure 5). To the point around differential expression, the benefit to random forest models is that they do not depend on individually differentially abundant bacteria, but take into account interactions and context of the features in a very robust way (e.g. non-linear relationships). Instead of thinking about the community in terms of statistically significant individual bacteria, we evaluated the signature across the whole community. In this way we were able to compare the overall predictive value of the virome in comparison to the bacterial communities. This is advantageous because, instead of thinking about the number of individual statistically significant values, we can observe more complex relationships between the disease and the total community. We added some more discussion around this point to the text. **LINES 120 - 123**

Additional comments: - A subset of virome samples had similar sequence depth as the negative controls. These were included in the analysis as the authors selected a low rarefaction depth of 1 million. This raises the concern that some of these samples may be more heavily influenced by environmental contamination. Did any of these low depth samples cluster with the negative control samples using Bray-Curtis distances? It would be helpful to show an ordination plot as part of Fig. S5 similar to that in Fig. S4A but with both true samples and negative controls to show that all true samples were distinct from negative controls.

This is a good point that we needed to clarify in the text. We omitted any samples below our rarefaction limit, which excluded most negative control samples and also a couple of the experimental samples. We had also included the ordination (shown as confidence intervals around the distances between centroids) to show that the negative control samples are very different from the experimental samples. **Figures S01, S05**

- Fig. S1B is mislabeled - sequence yield shouldn't be "ng/uL"

  We fixed this.

- It's not clear if the viral random forests classifiers for healthy vs. adenoma, adenoma vs. cancer, and healthy vs. cancer were significantly different in Fig. S7A as suggested by the text. Could the authors either annotate the figure to show significance, or adjust the text in the Results section to clarify that the trend was not significant?

Thank you for pointing out this omission. These differences were in fact statistically significant using a pairwise wilcox test and FDR corrected p-values. We updated the figure and associated text.

- The Results section referred to Figure 4E, but Figure 4 didn't have panels. Also, it's not clear why it wasn't Figure 3 (as the current Figure 3 wasn't cited until the following section).

  We fixed this.

- Fig. S9 should be incorporated into a main figure since it's the only evidence presented of phage-bacteria interactions potentially explaining the predictive power of viruses for colon cancer.

  We did this.