# Response to Reviewers

## Reviewer #1

**Several sections have been updated, and appropriately edited to more closely reflect the work done and the potential limitations.**

**I remain with the idea that the main weakness of the work is clearly that they recruited patients with various cancer type and various treatments, with then a quite low n in each subgroup. And that it would have been easy to clinically confirm their predictions for at least a two year-recurrence for the 26 patients included in the model discussed in lines 198-212. Data from these samples were previously published in Baxter et al Genome Med 2016, so I guess the sampling happens before 2015.**

We appreciate that our study does not have as many subjects as would be desired to reach stronger conclusions. It should be noted that we do not have the ability to control the frequency of people getting colon cancer, what drug regimen they are on, and so forth. We have done our best to pool patients by general categories and to not oversell what we are finding. Thankfully, although colon cancer is a leading type of cancer in the US, it is relatively rare and it is difficult to ask newly diagnosed patients to participate in such a study. We consider ourselves lucky to have the samples we have. To state the obvious, this was not an animal model study where we could control everything - this is human subjects research where there are clear restrictions placed on what we are able to do and who we are able to recruit. Related to this, because of IRB restrictions, we were not able to follow up with these patients to ascertain their current status (text added at L267). Regardless, we did not state that we could predict who would recur; we described how those treated for carcinomas had a microbiota that was largely similar to those with normal colons following treatment. Regardless, we would point out that Microbiome has published a number of high profile papers in the last year with similar or fewer number of subjects (e.g. dois: 10.1186/s40168-017-0328-9 [N=12], 10.1186/s40168-017-0305-3 [N=91 women], 10.1186/s40168-017-0309-z [N=10], 10.1186/s40168-016-0225-7 [N=18], 10.1186/s40168-017-0242-1 [N=80]). We have added clarification sentence in the methods study design and patient sampling section (L266-268).

**I have one minor point. Despite what is indicated in the response to the reviewers section, one of the comments has not been implemented in the text (last section of the third page of the reply to reviewers). Lines 67-74 state: "If the microbial community drives tumorigenesis then one would hypothesize that treatment to remove a lesion would affect the microbiota and risk of recurrence". As mentioned before, this sentence makes little sense to me, as it is already known that removing the lesion will affect the risk of recurrence, independently of a potential role for the microbiota.**

We have reworded this statement (L67-75) to make it more clear. Our point was that if there is something in the microbiota that promoted tumorigenesis then part of the therapy may also remove the bacteria that were driving the process and in turn reduce the risk of recurrence. Of course, removal of the lesion treats the disease, the question is whether a community that could cause recurrence would remain. Hopefully the reworded text have made this clear.

## Reviewer #2:

**While the revised manuscript is somewhat improved, there are still a few outstanding issues that I would argue need to be resolved:**

**(1) As of September 6, 2017, the link provided for analysis scripts (https://github.com/SchlossLab/Sze\_followUps\_2017 ) yielded a "404" not found error on github. I appreciate that this is just a broken link that can be easily fixed, but it made it more difficult to examine the authors' Random Forest approach.**

We apologize for the broken link. This has been corrected in the manuscript. The repository is available at https://github.com/SchlossLab/Sze\_FollowUps\_Microbiome\_2017.

**(2) I appreciate the authors detailed response on the machine learning questions. These are not easy question to resolve over review. I still am concerned, however, that every AUC they report appears to come from a model that "performs well" ( lines 121-122 and lines 133) even in cases (such as treatment for adenomas) where there is no other evidence that there was a difference between the conditions that the random forest model was trained on. What would set my mind at ease would be a demonstration that the authors' entire pipeline (including selection of the OTUs to model) produced an average AUC of 0.5 where there were no true differences between the groups (such as when labels are permuted).**

**The authors argue in their response: "The ranges reported in the text are the values of AUCs obtained after 100 different 80/20 splits of 10-fold cross validation. With the solid line in Figure S1 representing the AUC for the full set after 10-fold cross-validation. It is possible to do a permutation test but it would simply test whether our full set falls within the range of 100 different 80/20 splits of the same data.". I found this responses confusing. Isn't is possible to do 100 different 80/20 splits after permuting the labels on the data? Shouldn't repeats of the permuted data subjected to this process end up in a different range from 80/20 splits compared to unpermuted data? If that is not true, doesn't that suggest that the modeling has not produced any insights over the data (if the true labels for the data don't actually matter to the performance of the model)?**

**If the permutation of labels is repeated a number of times over the authors entire pipeline (including OTU selection and all parameter optimization), the average AUC should be 0.5. If the authors could provide such a result as a technical control it would demonstrate that an AUC range of, for example, 0.69-0.92, is the model "performing well", as they assert on line 121, and is not just a trivial reflection of an OTU selection procedure that biases the models to produce AUC's > 0.5. My sense is that the manuscript would be improved by such a demonstration. If this value on label permutations is in fact >0.5, it is reasonable to ask whether the model's performance is statistically different than one would expect over a dataset with permuted labels. If not, the authors should provide an alternative description of the model to "performing well" or at least indicate that the model would perform equally well on data with random labels and therefore biological insight in such cases is very limited.**

We appreciate the reviewer's concern that despite our best efforts our models are still overfit. To remedy this, we have used all possible OTUs to fit the model rather than the 10 most explanatory OTUs. Then by performing the label randomization, as suggested by the reviewer, we found that the models fit to those data indeed had an average AUC of approximately 0.5. When we re-assessed the AUCs for the correctly labelled models, we found that the pre vs. post-treatment models for the patients with adenomas and carcinomas were significantly better than the randomized data; however, the model for the patients with advanced-adenomas was not. The P-values that we include indicate the comparison of the AUCs for the models fit to the observed and randomized data (L126-127). We hope that this alleviates the reviewer's concern.

**(3) The conclusion that the cancer microbiome reverts to a more normal microbiome remains dependent on this complex machine learning pipeline with many tunable parameters that nonetheless only yielded a very modest p-value of 0.02. This is a limitation of the dataset, although to their credit the authors do not exaggerate the significance of their results.**

With the revised modeling approach, we actually observed that the P-value decreased. Regardless, the size of a P-value should not be used to assess the magnitude of the effect. Used as a threshold, the result was significant. Furthermore, as the reviewer points out, we did not go out of our way to exaggerate the significance of this result.