Response to Reviewers

The line numbers we refer to below are taken from the "clean" version of the manuscript we have submitted (manuscript.docx).

Reviewer #1:

Lesniak and colleagues colonize GF mice with different human fecal communities and challenge them with C. difficile RT027. They demonstrate an association between the initial microbiome composition and CDI severity. This study has advantages over prior work because the authors (1) use mice colonized with human microbiomes to maximize the microbiome beta diversity, and (2) control for factors that could contribute to CDI phenotypes such as host genetics and C. difficile strain. As a result, this study adds to a growing body of work that suggests the microbiome can affect susceptibility to C. diff infection.

The authors put quite a bit of attention to the supposed novelty of associating microbiome data with CDI severity, but the metrics for severity here (mortality, histopathology score, C.diff load, and toxin titer) seem to be tightly correlated. Previous studies have generally reported at least mortality and C. diff CFUs, so I'm not sure why this is new.

94 of the samples from Schubert et al. (the source of the fecal samples transferred to mice) were derived from patients who currently have CDI. Using these samples could potentially introduce C. diff from the stool samples instead of the initial challenge. Additionally, these microbiomes would already be greatly affected by C. diff and would not accurately reflect an initial microbiome that the authors are trying to transfer to the mice. Could the authors please clarify the samples used in Schubert et al. in a supplementary table? If they used CDI human samples, they may also want to reanalyze the data without these samples.

We have included a supplemental table, Table S1, with the donor labels and their associated metadata. Only one of the 15 samples came from a patient with CDI, donor M6. No C. difficile CFU was detected prior to day 1 in the mice that received stool from this donor. The human isolate from the experiments was the one cultured from this patient. We have clarified this in the Results (Lines 92-93) "After the gut communities had colonized for two weeks we confirmed them to be C. difficile negative by culture" and in the Methods (Lines 335-337) "Stool samples from each mouse were collected one day prior to C. difficile and plated for C. difficile enumeration to confirm no C. difficile was detected in stool prior to challenge." Lastly, the microbiome in the mice is not reflective of the human microbiome. The human microbiome is merely a source of variation. As when the human fecal community colonizes mice it shifts to be more similar to murine communities than human.

Detailed issues

Abstract/Introduction

Line 17-18: "Unclear whether there is an association between members of the gut microbiota and disease severity". Is this true? It seems clear by the following sentences that there is an association. Maybe be more specific. The manuscript "Gut microbiota-produced succinate promotes C. difficile infection after antibiotic treatment or motility disturbance" shows that succinate produced by B theta increases the severity of C. diff infection. Citation 15 also makes conclusions regarding the ability of gut bacteria to modulate sensitivity to C. diff infection.

We have edited this sentence for clarity to state we are investigating the associations between specific members of the gut microbiota and disease severity. (Lines 17-18) "unclear whether specific members of the gut microbiota associate with variation in disease severity"

End of Line 31: say C. difficile subtype

Change has been incorporated (Line 32)

Line 76-78: "We hypothesized that since specific groups of gut bacteria affect the metabolism of C. difficile and its infection dynamics, we can also identify groups of bacteria that affect the disease severity of the infection". Could the authors make it clearer what the difference between affecting infection dynamics and affecting severity of the infection is? Is this a difference between counting the number of C. diff that can affect, and the phenotype of the disease?

We have edited this sentence for clarity to better differentiate dynamics and severity. (Lines 76-79) "We hypothesized that since specific groups of gut bacteria affect the metabolism of C. difficile and its clearance rate, specific groups of bacteria associate with variation in CDI disease severity."

Could the authors specify more details about which samples from Schubert et al. 2014 (citation 17) were chosen and which mice the sample was given to. For example, maybe N1 was from someone who did not have C. diff but M6 was from someone who had C. diff or

antibiotics. The goal of the study is to see how the microbiome community affects disease phenotypes, but if samples are taken from people who have C. diff infection this could confound the study. Is it possible that the authors could be introducing C. diff from these fecal samples? Also, prior C. diff infections make subjects more susceptible to reinfection. As a result, these samples may not reflect what a susceptible microbiota may look like prior to C. diff infection.

We have included supplement table, Table S1, which provides information on each donor and referenced this table in the Methods (Line 335). However, the goal of this study it to investigate variation in the microbiome associates with differences in disease severity. We do not make any connection between these experiments and the human donors they came from since the community is more similar to murine than human, modifying any effect that patient susceptibilities may have had.

Figure 1 comments

Could the authors also show if the mice samples were more similar to the donor sample than to the other mice samples as well?

We have included the comparison between mice and their donor community in Figure 1B

Figure 2 comments

The observation that C. diff could colonize every mouse is surprising. It's possible fecal transfer resulted in diversity being lost, which could lead to susceptibility to C. diff infection. The authors may want to look at the C. diff colonization as a function of similarity to the original donor. Higher similarity to a healthy donor could result in less colonization.

At the time of C. difficile challenge, all mice have little similarity to their donor (added comparison to Figure 1B, IQR 0.9508-0.9891 median 0.9775). Furthermore when we tested for a correlation between CFU 1 dpi and community similarity to donor, by spearman's rank correlation, rho = 0.224 (p = 0.1178) and if we remove outliers below 0.9 (three samples from donor N1) rho = 0.0819 (p = 0.5844). Similarity to donor may contribute to the uncolonized N1 mice, however similarity to donor does not explain CFU for the remaining samples.

Figure 3 comments

Line 129: Could the authors specify which panel in figure 3 they are referring to when they say "Overall, there was greater toxin activity detected in the stool of the moribund mice".

Is this the summary panel? Could the authors provide the test used as well? Was linear regression used to correct for the donor? This does not seem to be specified in the methods

We used Wilcoxon test for differences in toxin activity and added an asterisk to Figure 3. We added a summary figure of toxin detection in mice prior to day 4 when there are no more samples for the moribund mice (Figure S1). We tested for significant difference in toxin detection between non-moribund and moribund mice with Pearson's Chi square test. We have added details to the figure legends (Lines 778-779, 813-817) "* indicates significant difference between non-moribund and moribund groups of mice by Wilcoxon test (P < 0.002).", results (Lines 130-131) "Overall, there was greater toxin activity detected in the stool of the moribund mice (Figure S1).", and methods (Line 389-391) "toxin detection in mice using the Pearson's Chi-square test".

Line: 132. Add that the observation seen is at timepoint 2. "Non-moribund mice from Donors N2 and N5 through N9 had comparable toxin activity as the moribund mice" only seems to be true in the middle panel of figure 3. By the end of the study, deceased mice have much more toxin for the most part.

We have clarified the timepoint of this observation. (Lines 132-134) "Non-moribund mice from Donors N2 and N5 through N9 had comparable toxin activity as the moribund mice at 2 days post-challenge."

Line 133: "Additionally, not all moribund mice had toxin activity detected in their stool". This statement is a little misleading. By the time the mice died, as show in in the bottom panel, all mice had very high levels of toxin activity detected in their stool. The authors can just add something like, "Additionally, not all moribund mice had toxin activity detected in their stool at every time point".

We thank the reviewers for bringing this unclear point to our attention. There were some moribund mice that never had toxin detected in their stool. It is unclear what leads this reviewer to the conclusion all mice had high levels of toxin activity. The bottom panel referenced could be either in reference to the bottom panel of Figure 3A or Figure 3B. Figure 3B is histopathological summary score, not toxin data. The moribund mice in the bottom panel of Figure 3A (10 days post-challenge) which has "Deceased" at 2.5 is noting that there was no toxin sample taken on that day since the mice were deceased. We have moved "Deceased" below the limit of detection line to avoid the chance the line could be interpreted at a toxin level for the deceased mice and added a note to the legend for Figure 2 and 3 (Lines 765-766, 779-780). "-Deceased- indicates mice were deceased at that time point so no sample was available."

Figure S1 comments

Could the authors also specify the test used to determine if moribund mice had higher histopathologic scores than the non-moribund mice?

We have clarified the statistical test (Wilcoxon) used in the figure legend of Figure 3 and S2 (Lines 778-779, 823-824) " indicates significant difference between non-moribund and moribund groups of mice by Wilcoxon test (P < 0.002)." and methods (Lines 387-388).*

Figure 4 comments

Generally, could the authors keep the timepoints used for bacterial abundances and tests used consistent throughout figure 4? Or provide equivalent figures using all types of analysis? Additionally, could the authors explain why correlations are used when looking at endpoint relative abundances but LEfSe used for timepoints 0? Doing this makes it difficult to compare results when different statistical tests are used for different timepoints.

We have switched all panels to use OTU level relative abundances, we adjusted the text to match (Lines 150-157). We used timepoint 0 since it is a common timepoint of challenge in order to compare non-moribund and moribund mice. There were no OTUs on day 0 that significantly correlated with histopathologic score after BH correction. We then looked at day 10 to test for the bacterial groups that still associated with histopathology. Since we are only looking at relative abundance and histopathologic score, we used a correlation instead of lefse. Lastly, we have added visualization of the high/low histopathologic score to figure 4C to better illustrate the split of the data in 4A to make the comparison easier.

Could the authors also say the timepoint the histology and toxin scores are derived from in panels A and B in the text? It appears that they are summary scores in the figure legend, but it is not entirely clear.

We have clarified the histology is from endpoint (Lines 149-150) "groups by severity level based on moribundity or 10 days post infection (dpi) histopathologic score for non-moribund", and toxin is the detection at any point in the experiment (Lines 158-159) "explain the differences between communities that had toxin activity detected at anytime point to those that did not".

Could the authors explain why they chose to make histological score and toxin activity into binary variables instead of a continuous one for panels A and B and use continuous data in panel C?. It makes it harder to compare the influence of Bacteroides in panel A and C when the type of outcome variable and test used are different. Are the cutoffs clinically relevant?

We made toxin binary (detected or not at any time point), because we sought to identify associations with probability of toxin production and not the level of toxin activity produced. For histology, to incorporate the more severe infections that resulted in moribundity which had their histology scored at a different time point in the infection (peak infection) it cannot be directly compared to values of the endpoint histopathology scores. So we instead looked at overall outcome grouped into three categories of low to high severity of non-moribund with low histopathology (<5), non-moribund with high histopathology (>5), and moribund. Furthermore, the non-moribund histopathology scores had a bimodal distribution (Hartigans' dip test, p < 0.05, added clarification to text (Lines 374-377, 388-389). "For non-moribund mice, histopathological summary scores used for LEfSe and logistic regression were split into high and low groups based on greater or less than the median summary score of 5 because the had a bimodal distribution (P < 0.05)." "non-unimodality to non-moribund histopathological summary score using Hartigans' dip test"
Finally to aid in comparing panels A and C we have added color to Figure 4C to show which points low/high histopathology scores (Lines 792-794). "Each individual mouse is plotted and colored according to their categorization in panel A. Points at the median score of 5 (gray points) were not included in panel A."

In addition to doing correlations in Figure 4C which don't account for factors like donor group could the authors used a mixed effect model as well where they keep histology score a continuous variable? LEfSe the software package used to do the analysis should also be able to accommodate this.

Since the variation in the community is driven by the donor, donor would be a fixed variable and we would not want to remove this effect. We tested for effect of donor and community on summary scores with PERMANOVA, which both had a significant effect (P < 0.01). We are unaware of a LEfSe package able to handle continuous outcome.

Could the authors include p-values in figure 4? Could the authors also order the plots by fold change or p-value instead of alphabetical order? This may help make it easier to see which microbes change the most between groups. Using alphabetical order to ease comparisons is a good goal. If alphabetical order is desired, it may be useful to include the union of the significant microbes in panels A and B and make them similar sizes to allow for direct comparison between the 2 figures.

We have sorted by LDA and included LDA and p-values to the Figure 4A and B.

In lines 167 to 169 the authors specify that Kelbsiella and Prevotellaceae were positively correlated with the histopathologic score and increased in the groups of mice with detectable toxin. Could the authors specify that this observation is seen at 2 different time points (timepoint 0 and the endpoint)?

Updating Figure 4 panels A and B to use OTU has removed these genera from significance. Thus, we have removed the text adjusted the text to match the updated observation (Lines 167-168). "A population of Bacteroides, OTU 17, was positively correlated with the histopathologic score and were increased in the group of mice with detectable toxin."

Could the authors do a different version of Figure 4C but do toxin instead of histology score?

We tested for a correlation between toxin activity level and OTUs, but none were significant (P < 0.05 after BH correction). We added this clarification "We also tested for correlations between the endpoint relative abundances of OTUs and toxin activity but none were significant." (Lines 168-170).

Figure S2 and 5

Line 178 to 180: "C. difficile was more likely to produce detectable toxin when the community infected had less abundant populations of Verrucomicrobia and Campilobacterota" may too strong of a statement. As shown in figure 5A Verrucomicrobia have almost the same mean abundance in subjects with and without toxin. Is this formally tested or just assumed based on the mean abundance? If you would like to formally evaluate direction of association from the Random Forest model the authors can use an R package called LIME (https://www.analyticsvidhya.com/blog/2021/01/ml-interpretability-using-lime-in-r/) or shapely values (https://cran.r-project.org/web/packages/shapr/vignettes/understanding_shapr.html). The authors could also use this method to confirm the statement "The features with the greatest effect showed that communities with greater populations of bacteria belonging to Bacilli and Firmicutes and reduced populations of Erysipelotrichia were more likely to result in moribundity" as well as the equivalent statement in line 194 for histopathologic scores. Alternatively, the authors could use logistic regression and report if there is a significant increase or decrease in odds of toxicity with an increase in abundance of a microbe.

We understand the complexity to the random forest model and issue with interpreting features, so we have switched the model to use logistic regression and reported the odds ratio for

a more interpretable feature value. We have updated the bacteria members of importance in the text (Lines 176-190, 393-399). "We built L2 logistic regression models using the mikropml package (87). Sequence counts were summed by taxonomic ranks from day 0 samples, normalized by centering to the feature mean and scaling by the standard deviation, and features positively or negatively correlated were collapsed into a single feature. For each L2 logistic regression model, we ran 100 random iterations using values of 1e-0, 1e1, 1e2, 2e2, 3e2, 4e2, 5e2, 6e2, 7e2, 8e2, 9e2, 1e3, 1e4 for the L2 regularization penalty with a split of 80% of the data for training and 20% of the data for testing."

Could the authors comment on why there are no error bars in the Histopathologic score panel in Figure S2? It would also be helpful for evaluation of the models to put dots for the AUC of each model fit instead of just the IQR and mean.

Since changing to logistic regression, the plot in question for Figure S2 with no error bars has been removed. We have added individual points for AUC in addition to the mean/IQR.

Figure S2 legend: Could the authors clarify the statement "Individual relative abundances were added to F since differences in AUC were outside the interquartile range". Could the authors also clarify what the grey dots are in Figure S2, panel F? Could the authors also provide an explanation for why there are no error bars in in figure S2, panel F?

Since changing to logistic regression, Figure S2 panel F and the statement has been removed.

The authors interpret figure S2, panel F by saying, "No genera had a significantly greater effect on the model performance than any other, indicating the model was reliant on many genera for the correct prediction" An alternate, but complementary possibility is that many of the microbiome features could be co-linear. Could the authors produce a heatmap on a correlation matrix of microbial features to check this? If there are features that have high collinearity (>0.9) could they remove highly co-linear features and fit the model again? Co-linearity should not affect prediction, but it could affect the feature importance.

We have accounted for co-linearity in feature importance. Correlated features already been removed. In the preprocessing step of the mikropml package, features positively or negatively correlated featues are collapsed into a single feature. We have included these details of the mikropml preprocessing step in the methods "Sequence counts were summed by taxonomic ranks from day 0 samples, normalized by centering to the feature mean and scaling by the standard deviation, and features positively or negatively correlated were collapsed into a single feature." (Lines 394-396)

Discussion

Line 264-266: "Akkermansia was most abundant in the non-moribund mice with low histopathologic scores but there were some moribund mice which had increased populations of Akkermansia" could be slightly misleading. This is because, on average, Akkermansia is more abundant in moribund mice even though the maximum abundance of Akkermansia is higher in non-moribund mice. Could the authors change the sentence to specify this?

Thank you for bringing this potentially misleading statement to our attention. We understand how it is unclear and could be misinterpreted. We have updated this statement with the OTU results "In our data, a population of Akkermansia, OTU 5, was most abundant in the non-moribund mice with low histopathologic scores but moribund mice had increased population of Akkermansia, OTU 8." (Lines 257-259)

Reviewer #2:

The article "The gut bacterial community potentiates Clostridioides difficile infection severity." by Lesniak et al. describes a mouse model of CDI. Previously germ-free mice were colonized with different microbial consortia in the form of homogenized stool samples from different donors. Two weeks after this inoculation, mice were infected with identical spore counts of a single C. difficile strain. Stool samples were collected on the day of C. difficile infection and the following days, and C. difficile CFU counts were obtained on day 1 and 10 (or at death).

The paper addresses a research question with insufficient evidence in the literature: does the microbiome modulate CDI severity? I find the experimental set up to be convincing and appropriate, and the data collected to be useful but limited. Crucial aspects relevant to the research question cannot be addressed, and the analyses performed on the available data are not optimal. Therefore, I am not convinced by several major results. I would ask the authors to perform several additional analyses on their data, and at least comment on the issues that cannot be addressed by their data as collected.

The unaddressed questions that cannot be addressed with the data collected: What is the effect on the previously germ free host of the different microbial communities? The introduction states "Since the murine host and C. difficile isolate were the same and only the gut community varied, the variation in disease severity we observed was attributable to the gut microbiome." This ignores the potentially drastic differences in host responses to the different inocula. Since the paper's main conclusions rest upon the d0 microbiota

differences as predictors of responses, this ignored host response aspect must at least be discussed.

We thank the reviewers for bringing this issue to our attention. We have included discussion about the different responses the host can have in response to different inocula "Different gut microbial communities can also have different effects on the host immune responses" (Lines 298-299).

Unaddressed questions that could be addressed: Longitudinal fecal samples were collected, but no longitudinal microbiota data are presented. Previous research has shown multiple important ecological interaction between gut bacteria and C. difficile, including for example between C. scindens and C. difficile (PMID: 25337874). This work from 2015 shows a mechanism of ecological interference but is not cited. Do the researchers here find different/consistent dynamics of C. difficile over time across mouse groups? Could those dynamics help explain different outcomes? Since the paper's findings generally come from using d0 samples to make predictions of outcomes, would an analysis of the C. difficile trajectories not enrich the insights? I am not sure why the authors did not do these analyses since they do assert that all mice could be infected (with the exception of two mice from donor n1). What happens next to the C. difficile populations? We are not told except that at d10, higher CFU counts are observed than at d1.

*We thank the reviewer for identifying this unaddressed question. While it is an interesting question, we found the dynamics to be fairly linear and agree with the reported results. We have added Figure S4 and to the methods and discussion to state "These bacteria groups increased in severe outcomes maintained their differences throughout the length of the experiment (Figure S4)."(Lines 282-283, 383-386, 837-844) "We tested for differences in temporal trends through fitting a linear model to each OTU and testing for differences between histopathological summary scores with LEfSe (85) in mothur (default parameters, LDA > 3)." "**Figure S4. Temporal dynamics of OTUs** that differed between histopathologic summary score. Relative abundance of OTUs on each day relative to the time of C. difficile challenge (Day 0) that have a significantly different temporal trend by the histopathologic summary score by LEfSe analysis. Median (points) and interquartile range (lines) are plotted. Points and lines are colored by infection outcome of moribund (colored black) or non-moribund with either a high histopathologic score (score greater than the median score of 5, colored green) or a low histopathologic summary score (score less than the median score of 5, colored light green)."*

The microbiota analyses are not clear:

- it is very interesting indeed that the individual histopathologic scores differ between mouse/donor groups. However, N7 mice appear to score low on the individual scores yet have a high summary score. I do not understand this. the numbers shown in figure s1 seem to not add up to the numbers in figure 3.

We thank the reviewers for calling attention to this discrepancy. There had been a labeling error in the supplemental figure S1 (now figure S2). It has been updated to match Figure 3.

- regarding bacterial abundance analyses and outcomes performed with lefse: did you first contrast moribund vs non-moribund (e.g. all other mice) , and then dichotomized among non-moribund mice? This should be explained better.

This is correct. Mice are either moribund or non moribund, and the non-moribund are dichotomized. We have edited the text to clarify this categorization. (Lines 149-150) "We split the mice into groups by severity level based on moribundity or 10 days post infection (dpi) histopathologic score for non-moribund."

- please perform a regression analysis of microbiome features on disease scores instead of / in addition to a lefse performed on the somewhat arbitrarily dichotomized disease severity score. you could do this on each histopathology observation individually, or more elegantly, by a model that recognizes the histopathology categories and integrates predictors across them. Do some bacterial communities prevent specific types of damage? do the same bacteria prevent all damage types? You have continuous/ordered categorical outcomes, why focus only on classification?

Disease severity score was dichotomoized because it had a bimodal distribution (Hartigans' dip test, $p < 0.05$). We have added clarification to text (Lines 374-377, 388-389). "For non-moribund mice, histopathological summary scores used for LEfSe and logistic regression were split into high and low groups based on greater or less than the median summary score of 5 because the had a bimodal distribution ($P < 0.05$)." "non-unimodality to non-moribund histopathological summary score using Hartigans' dip test"

In our initial analysis, the individual histopathological score components did not show significant differences in predictors than those of the summary scores.

- In general, the random forest classifier models are not explained sufficiently. "To determine the optimal split, we tested splits (50%, 60%, 70%, 80%, 90% data used for training) to find the greatest portion of data that could be used to train the model while still

maintaining the same performance for the training model as the model with the held-out test data." The second part of this sentence, "as the model with the held-out test data" is unclear to me. What are you optimizing here and why? If the same mice were held out to perform what this sentence is suggesting, this sounds like an overfitting exercise to me. Furthermore, were features transformed from relative abundances, or standardized in any way?

To use a more interpretable model, we switch to using logistic regression instead of random forest and removed different data splits and the unclear explanation as well. Lastly, in the preprocessing step of the mikropml package, features are normalized by centering on the mean value and scaled by the standard deviation. This information has been added to the methods (Lines 393-399)." We built L2 logistic regression models using the mikropml package (87). Sequence counts were summed by taxonomic ranks from day 0 samples, normalized by centering to the feature mean and scaling by the standard deviation, and features positively or negatively correlated were collapsed into a single feature. For each L2 logistic regression model, we ran 100 random iterations using values of 1e-0, 1e1, 1e2, 2e2, 3e2, 4e2, 5e2, 6e2, 7e2, 8e2, 9e2, 1e3, 1e4 for the L2 regularization penalty with a split of 80% of the data for training and 20% of the data for testing."

- Because of this, I worry about the correctness of the results. Laudably, the authors go back to raw data following their complicated models. But in line 180 Verrucomicrobia are highlighted as associated with less toxin production, yet figure 5a shows no striking differences. Worse still, for Campilobacterota, the opposite trend from the narrative in the manuscript/model appears in the raw data. Please discuss/explain.

We have removed these references based on the random forest models and switched to reporting the odds ratio from the logistic regression for more interpretable features.

- Figure 5: microbiota predictiveness of toxin production is performed at the phylum level, moribundity by class, and disease score by genus as a consequence of the model choice algorithm. This means however, if we look at figure 5a vs c, that for toxin production, close to 100% of the community was considered in the analysis, while for the disease score analysis, the main focus of the paper, much less of the community was considered (e.g. generously considering the relative abundances in 5c to be 3% each, this would sum to at most 45% of the community. More likely, less than 10% of the community was considered in the disease score analysis. Most importantly, the chosen genera include surprisingly few Clostridia genera, but the authors state...

- ....Line 190: "No genera had a significantly greater effect on the model performance than any others" It is not clear how these genera were chosen, and from what original list of genera. None of them are very abundant, whereas we can see in figure 5b that clostridia

make up over 95% of the non-moribund population. Where are the missing Clostridia genera? Could this be an artifact of the random forest classifier that identified one genus of Clostridia as similarly predictive as another, therefore choosing only one in the consensus tree? Does this really mean the left out genera would have no biological role in CDI severity modulation? I suspect that much of the variation in relative abundances of the very low abundant taxa was driven by differences among the dominant Firmicutes/Clostridia. This needs to be better discussed or analyzed. As a practical suggestion, I would recommend a logistic regression (if the authors insist on a binary outcome) with genera/microbiota features selected based upon prevalence considerations, or better still an ordered categorical regression to relate microbiota features to outcomes on an ordered scale.

To clarify, in the previous version of this figure being referenced here only the features with an median decrease in performance when permuted > 0 were being plotted. The models were using the information from the whole community in all models. However, we have since switched from random forest to logistic regression and the classification level used. In these models, all features are used, but we are only plotting features with a mean odds ratio not equal to 1. We have clarified this in the figure legend (Lines 802-803). "Models used all community members but plotted are those members with a mean odds ratio not equal to 1."

- In the discussion, we are presented with a new result: "In our data, Akkermansia was most abundant in the non-moribund mice with low histopathologic scores but there were some moribund mice which had increased populations of Akkermansia", inspecting figure 4a suggest a different finding, that moribund mice on average had the largest Akkermansia levels.

We thank the reviewers for bring our attention to this unclear statement. This has been adjusted based on the update of Figure 4A to OTU level and clarified the statement further (Lines 257-259). "In our data, a population of Akkermansia, OTU 5, was most abundant in the non-moribund mice with low histopathologic scores but moribund mice had increased population of Akkermansia, OTU 8."

Overall, I think the authors have conducted a good experiment. The unknown physiological effects of colonization by different microbiota on the germ free host ought to be discussed. The analyses should be revisited to ensure that the major findings pertaining to the function of specific taxa in CDI severity modulation hold up.

*** minor
- line 60ff: not clear what "barrier" means here. please be more precise or define it. competitive exclusion? physical barrier?

We have removed ambiguous "barrier" from Line 61 and replaced "barrier" with "inhibition" in Lines 62 and 64.

- line 60: "...preventing C. difficile from infecting the gut." This statement should probably provide the most prominent reference for this claim, Buffie et al. 2015 Nature. PMID: 25337874

We have added the suggested reference (Line 61).

- line 66: again, missing PMID: 25337874 citation

We have added the suggested reference (Line 66).

- line 91: show they equilibrate? if you do not show it, how can I know it happened? you could instead say that you waited for two weeks.

We have changed equilibrate to colonize (Line 92).

- thetayc beta diversity analyses should be verified on less stringently rarified data. 2107 sequences (reads?) per sample is very low, and given the dominance of clostridia shown in figure 1 may suggest that the differences in very low abundance taxa are accentuated by this cut off.

To reduce the risk of low abundant taxa affecting thetayc, the rarification performed sub-samples the data 1000 times and averages those 1000 sub-samplings. We have clarified this in our methods (Lines 379-380). "For this calculation, we averaged of 1000 sub-samples, or rarified, samples to 2,107"

- please provide evidence/state that the donor samples and the mice prior to CDI infection were free of C. diff.

We have added to the methods the statement "Stool samples from each mouse were collected one day prior to C. difficile and plated for C. difficile enumeration to confirm no C. difficile was detected in stool prior to challenge" (Line 337-339).

- what lefse parameters were chosen?

We have added the requested details to the methods. "default parameters, LDA > 4" (Line 382).

- line 164: what are "endpoint relative abundances" d10 samples/at death samples? line 165: the sentence is difficult to parse.

We have clarified these are "day 10 relative abundances" and added "(10 dpi)" to the axis label as well. (Line 163).