**Reviewer #1**

- Concerns about novelty/significance - All the described microbial associations with CRC / adv. adenomas merely confirm what has been published before, novel associations are not presented (as the authors acknowledge).

The primary goal of this study was not to identify bacteria associated with CRC, as that has been done extensively (though not in a cohort as large as this one). Our focus was testing whether such associations are useful when used in combination with FIT, the current state-of-the-art non-invasive screening method. Nonetheless, we assert that confirmation of known associations in this large cohort from four different study sites makes a significant contribution to the field.

- The paper essentially presents the negative finding that the microbiome does not allow for more accurate detection of cancer or adv. adenomas compared to the FIT, in particular for high specificity. The presented combination of the microbiota-based test with the FIT does not substantially improve performance over FIT alone for detection with high specificity.

While we appreciate the concern for high specificity, the purpose of CRC screening is not to prevent false positives. If that were the case, one could argue to forgo any screening. No screening would eliminate all false positives. The intent of CRC screening is to find asymptomatic individuals with adenomas. The primary purpose is sensitivity – finding true positives, with an important secondary objective of minimizing false positives. In the current national climate of apprehension regarding unnecessary testing, we appreciate this concern but maintain that information from the MMT can offer additional options for both research and clinical practice.

Please keep in mind that the purpose of screening is not to diagnose cancer. When cancer has already developed, the likely benefits of population screening are tenuous (options for treatment are narrower and differences in long-term survival less evident). Thus, the statement that the microbiome does not allow for more accurate detection of cancer does not speak to the underlying goal here. The goal of CRC screening is to find early stage cancers, pre-invasive lesions, and *adenomas* in asymptomatic people – and the MMT did this better than FIT.

In summary, we disagree with Reviewer #1 that this is essentially a negative study. The MMT outperformed FIT in detecting adenomas. Beyond this, the MMT furnishes information regarding the microbial environment in which cancerous and noncancerous lesions develop. This can form the basis of studies that may eventually prevent colorectal cancer.

Comparisons of sensitivities at differing specificity cutoffs are difficult and the conclusions drawn are often misleading or wrong. This affects the Abstract, which only mentions sensitivities, but not specificities, clearly misleading the reader. Moreover, the corresponding Results sections (starting with line 190) and Figs 1D, 2BC, 3 and 4, I consider uninformative/inconclusive because of this issue. Comparing positive predictions and sensitivity clearly makes most sense if the specificities are held constant. The authors could easily fix this by choosing one or more consistent specificity cutoff (between 0.9 and 1, see below) and compare the respective sensitivities for all methods tested. The R package pROC provides a statistical framework to establish whether sensitivity significantly differs at a given specificity, a much more meaningful test than McNemar's method. Because of this flaw I feel that the authors' conclusions about increased sensitivity / detection rates of their microbiota-based test (or the MMT, which is its combinations with FIT) are entirely unjustified.

With initial screening studies that investigate factors on a continuum, it is important to present ROC curves (giving the results across the entire spectrum of sensitivities/specificities) as well as the probabilities of lesions and FIT results (across the entire range of values). These are given in the figures.

Most tests proposed previously for CRC screening have higher sensitivities than the microbiota-based tests proposed here (including the MMT). This is because CRC incidence is low (the authors state 0.3% themselves), meaning that if 1000 individuals are screened, there will be <= 3 true positives, but even at a specificity of 95% about 50 false positives. So, although I am not a clinician, I am convinced that the potential of any screening test with a specificity <90% is very limited, because of the excessive costs (and other harm) it would incur for re- examining the many false positives by colonoscopy to arrive at a correct diagnosis. In this high-specificity regime, MMT does not provide an advantage over FIT alone (Fig 2A).

We take issue with the statement that most tests proposed previously for CRC screening have higher sensitivities than the microbiota-based test proposed here. In a recent meta-analysis of accuracy studies (*Ann Intern Med*. 2014 Feb 4;160(3):171), the pooled sensitivity of FITs was 79% (for colorectal cancer) with a positive predictive value of 13%. The United States Preventive Services Task Force indicates that the sensitivities for various FITs range from 61% to 91% for colorectal cancer (Screening for Colorectal Cancer: An Updated Systematic Review, AHRQ Publication No. 08-05-05124-EF-1). There are fewer studies which report the results for detecting adenomas (the specific target in public health screening). In the Imperiale et al report (*N Engl J Med* 2014;370:1287-97), the multitarget DNA test had a sensitivity of 17.2% for non-advanced adenoma and 42.4% for advanced precancerous lesions. The FIT test had a sensitivity of 7.6% for non-advanced adenomas and 23.8% for advanced precancerous lesions. In our study, the sensitivity of the MMT was 57.1% for adenomas and 95% for cancerous lesions.

The reviewer states, “This is because CRC incidence is low (the authors state 0.3% themselves), meaning that if 1000 individuals are screened, there will be <=3 true positives.” The reviewer is confusing incidence with positive predictive value. These measures are related but are not equivalent. “If 1000 individuals are screened, there will be xx true positives” is a statement regarding positive predictive value. Positive predictive value is dependent upon the prevalence of the disease, but also the sensitivity and specificity.

The NCI-funded Early Detection Research Network has supported several large discovery and validation studies of non-invasive screening tests for colorectal cancer. This consortium of experts has established two criteria to move biomarkers from discovery to validation. After discovery, a biomarker’s initial screening performance is assessed using convenience samples, that is, samples obtained from non-controlled environments that are not well annotated. For entry into a reference set, the experts set a cut off performance in a convenience set of 50% or higher true positive rate and 30% or lower false negative rate. They set these cuts offs purposefully low to include any biomarker that might have potential as a screening marker. In the recent NCI supported validation study of new non-invasive screening tests for colon cancer supported by NCI (U01 CA86400), the experts and reviewers designed a study to the performance of Test A, a binary test, to FIT. To make a fair comparison, they lowered the threshold of FIT to the specificity corresponding to that of Test A (preliminary data suggested to be around 83%). Sensitivity at this lower threshold will be at least the same or higher than that based on the threshold of 100 ng/ml for FIT. Therefore, it is not uncommon to set generous thresholds in the preliminary stages of investigation.

We agree that there are costs and harms associated with unnecessary testing due to false positives. But, there are also costs and harms associated with not detecting cancer in people who truly have the disease. These people should also not be neglected. With knowledge gained from both FIT and MMT, we argue that this may eventually provide additional options for physicians and patients. It also recognizes that more data (whether the types of bacteria in the colon, the age of the patient, the genetic profile of the patient, etc.) have the potential to enhance prediction. More data – not less – will shape the future of precision medicine.

Hardcastle, J. D., J. O. Chamberlain, M. H. Robinson, S. M. Moss, S. S. Amar, T. W. Balfour, P. D. James and C. M. Mangham (1996). "Randomised controlled trial of faecal-occult-blood screening for colorectal cancer." Lancet **348**(9040): 1472-1477.

Jorgensen, O. D., O. Kronborg and C. Fenger (2002). "A randomised study of screening for colorectal cancer using faecal occult blood testing: results after 13 years and seven biennial screening rounds." Gut **50**(1): 29-32.

Mandel, J. S., J. H. Bond, T. R. Church, D. C. Snover, G. M. Bradley, L. M. Schuman and F. Ederer (1993). "Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study." N Engl J Med **328**(19): 1365-1371.

Scholefield, J. H., S. M. Moss, C. M. Mangham, D. K. Whynes and J. D. Hardcastle (2012). "Nottingham trial of faecal occult blood testing for colorectal cancer: a 20-year follow-up." Gut **61**(7): 1036-1040.

Statements that one method is more accurate than another should be supported by statistical tests (but are lacking in Fig 2A).

We did not make any statements that one method was more accurate than another. We made comparisons of the AUC (Fig 2A) and sensitivity (Fig 3, Table 1). The statistical tests used are described in the methods section, and all p-values are presented in the text.

The authors' reasoning on the serial combination of the FIT and their combination method (MMT) in Results and Discussion is flawed. If the FIT is applied first, a serial combination test would only maintain its specificity, if the negative results from the initial FIT were NOT retested. Retesting these negatives with MMT as proposed by the authors will not only result in additional true positives, but also in more false positives; in fact, the proposed combination will have a HIGHER false positive rate THAN ANY of the individual methods, because false positives are accumulated from both tests. There is no free lunch in statistics either: if a strategy like the proposed retesting of FIT results with a method that is itself a combination with the FIT test worked like the authors suggest, one could imagine recursively applying this approach to reduce the error rates to arbitrarily small values, which is of course impossible. Proper evaluation of serial testing would have made these issues apparent.

The reviewer’s comments regarding serial testing are well taken. Therefore we have removed the serial testing from the results section and instead focused on estimates of trues positives identified by FIT or MMT alone. We revised Table 2 to contain only projected population figures from FIT testing alone and MMT testing alone.

The Methods description does not clearly state whether random forest classifiers were cross-validated. Using the built-in out-of-bag (OOB) error estimate of the random forest classifier is generally not accepted as accurate evaluation of its accuracy (although I agree with the authors that it is a nice property of the RF classifier that helps preventing extreme overfitting).

We disagree with the reviewer that the OOB error estimate of the random forest classifier is generally not accepted as accurate evaluation of its accuracy. The OOB error estimate has been repeatedly shown to more accurately reflect the generalization error than using cross-validation, especially when the number of samples is larger than the number of features, like the models in this study. Splitting the data with an additional cross validation does not improve your confidence in how well the models perform, because either way, the models are only tested on data they haven’t seen. Furthermore, instead of training on 2/3 and testing 1/3 for each tree (i.e. the OOB error rate), you would be training on ~1/2 of the data, ignoring ~1/4 (the OOB error), and basing your findings on the remaining ~1/4 that you set aside as a training set. This reduces power and gives a less accurate estimate of the real error rate. Nonetheless, as other reviewers have pointed out, the MMT model needs to be applied to an independent validation cohort to more accurately assess its performance. This is a future goal of our research group.

Breiman, L. (1996). *Out-of-bag estimation* (pp. 1-13). Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708, 1996b. 33, 34.

Mitchell, M. W. (2011). Bias of the Random Forest out-of-bag (OOB) error for certain input parameters. *Open Journal of Statistics*, *1*(03), 205.

The Discussion on a potential confounder that is due to taking stool samples after colonoscopy is very inaccurate. The main cohort in Zeller et al. MSB 2014 was sampled BEFORE colonoscopy (as stated in their Methods); also the cohort described in Feng et al. Nature Communications 2015 was sampled before colonoscopy (personal communication with the authors, not obvious from the publication). Yu et al. Gut 2015 partially sampled before colonoscopy; these authors corrected for this potential confounder in their statistical data analysis.

We thank the reviewer for pointing out this error. It has been corrected in the text.

Feng et al. Nature Communications, March 2015 is a relevant study which should not be omitted here

A citation for this study has been added to the introduction.

**Reviewer #2**

- One question I think that readers will have is how robust these results are to different sequencing pipelines. It would be of interest whether this dataset could be used as a training set to construct a machine-learning model that could then make case-control assignments with sequence data generated from one or more of the previous studies the authors cite. (This would have to be done at the genus level if previous studies did not share an overlapping region of the 16S rRNA gene) If the models failed to make these predictions, it would suggest that the changes associated with lesions were of smaller magnitude than variation associated with different pipelines. I suspect this might be the result and would suggest that future commercialization of this process will require substantial effort to control this sort of variation.

We would very much like to apply the MMT to an independent dataset. However doing so would require 16S rRNA gene sequencing data **with matching FIT results**. To our knowledge, such a dataset is not yet available. Nonetheless the bacterial populations associated with CRC in our dataset have been observed by other groups using different sequencing pipelines, suggesting these associations are robust to sequencing method. Furthermore, we acknowledge that the application of 16S rRNA gene sequencing to human samples is still an active area of methods development. Myriad variables ranging from sample storage, DNA extraction methods, variable regions, primers, sequencing platform, etc. may have an impact on the results. As indicated to Reviewer #1, our next step will be to validate this model with an even larger cohort. That there is overlap in the types of bacteria that enter into our model and previous efforts should be considered a strength.

Are the sequences and scripts used to generate the model freely available to allow other investigators to reproduce the figures and explore construction of alternative machine learning models?

Yes, raw data and scripts are publicly available via the NCBI SRA and GitHub. See ‘Availability of data and materials’ section (line 352).

**Reviewer #5**

A weakness is that the results are not further tested in a separate, verified set of samples for reproducibility in a different population. Thus, the generalizability of the results remains to be determined.

We agree with this reviewer and the previous comments and are working towards acquiring and analyzing samples from a validation cohort. As mentioned above, applying the MMT to another dataset would require matching FIT samples, which is not available in the literature.

The presentation of the results would be easier for the reader to understand if presented in a more parallel fashion. For the initial comparison, FIT is compared to microbiota to distinguish healthy from those with adenomas yielding a model using 22 bacterial populations. In contrast to later sections of the paper, no specific species contributing to this model are mentioned in the main text. Some description in the text would be helpful to the reader beyond Figure S1A.

We have added a general description of the taxa in the normal vs. adenoma model microbiota model to make the presentation more parallel (line 154).

This approach is very similar to the 2014 Zackular et al. study (from the same laboratory) which also produced impressive results using  a series of logit regression models. For example, the Zackular study found that age, gender and race appeared to enhance separation of carcinoma and normal samples. Did the authors consider this type of patient metadata in their study? The authors did evaluate MMT with these factors, but would they enhance the results in Figure 1?

We tested a model combining the MMT with patient metadata. We have added a paragraph to the results section and a supplemental figure with the result of this model (line 247).

Given that the authors had the Zackular et al. models on hand, did they evaluate the performance of those models on this new and much larger dataset? Both datasets appear to utilize the same primer set and sequencing technology.

We chose not to include the Zackular models in this study. First, it became clear to us that the logit models that we developed in the Zackular et al. study were likely overfit to the data. This was why we decided to apply the Random Forest algorithm because it is widely held as being robust to possible overfitting. Second, the Zackular et al. study focused primarily on models that used 16S rRNA gene sequence data. In the current study we combined the microbial data with the FIT results. The earlier models did not use FIT. We were pleased that the overall compositions of the models overlapped.

- The primary concern with the modeling approach is the possibility of over-fitting RF models which is known to occur despite efforts to perform internal cross-validation. Why didn't the authors utilize the Zackular et al. V4 stool-based CRC dataset for external validation? This dataset had 30 CRC, 30 Adenomas and 30 normal samples. This would substantially add to the models evaluation and lend support to use of the word "predictive" throughout the text. It would also help to establish whether the current RF methodology may be over-fitting.

As discussed above, the machine language literature and the consensus of experts that we have talked with indicates that Random Forest is not sensitive to overfitting. The stool samples used in the Zackular study were a subset of the samples used for this study. By removing them from the current study to develop a training set and then testing the model against the Zackular data we would be arbitrarily dividing the dataset and hurting our modeling power in the process.

The reader is struck in reviewing the individual comparisons of normal vs adenoma or normal vs CRC that there is limited overlap in the OTUs. What is the hypothesis for this result?

We hypothesize that the loss of beneficial organisms or their metabolites could promote inflammation that facilitates the transition from adenoma to carcinoma, while enriching for organisms, such as Fusobacterium or Porphyromonas, that are better adapted to live in an inflamed environment. Based on our results, it seems that the loss of potentially beneficial organisms is a better predictor for identifying adenomas, while the emergence of potentially pathogenic organisms is more predictive for carcinomas.

It is also concerning that the 34 OTUs in the microbiota carcinoma model from Fig S2 and the 23 OTUs in the MMT model from Figure S3 share very few OTUs -- only six OTUs appear to be shared. Could the authors comment on this shift in signal in the discussion? Are the original OTUs in the carcinoma model that drop out in the MMT model highly correlated with FIT quantitative measurements? Were the new OTUs used in the MMT model marginally significant in the first carcinoma model?

As the reviewer predicted, many of the OTUs that drop out from the microbiota carcinoma model to the MMT model are significantly correlated with FIT, effectively decreasing their importance when used in combination with FIT. Instead, the MMT is enriched with OTUs from the adenoma microbiota model (13 OTUs in common). Since FIT is sufficient to detect most carcinomas, the MMT leverages those adenoma-associated OTUs to improve the sensitivity for adenomas that are undetectable by FIT alone. We have expanded upon this topic in the discussion (line 297).

The Discussion section should include a paragraph on the study and method limitations as well as the authors' vision for 'next steps'. It doesn't seem that either the sensitivity or specificity of the test (nor the degree of data overlap among the groups as shown in Figures 2 and 4) would be dissuasive of the need for colonoscopy but a perspective from the authors on clinical use would be helpful to the reader.

We have added a paragraph addressing some of the limitations of this study to the discussion(line 322), including the fact that the MMT could not be applied to a separate validation set. We also added a brief perspective on the potential clinical use of the MMT. (line 268)

**Minor Concerns:**

We have also addressed each of the following minor concerns raised by the reviewers.

- It seems line 156 should say 15.7% of adenomas not cancers based on the rest of the paragraph.

- line 186 (ref 17 should also be included here).

- Line 79: "In the present study demonstrate the potential for microbiota" should read "In the present study, we demonstrate"

- Line 161: "with cancer using the relative abundance 34 bacterial populations" should read "with cancer using the relative abundance of 34 bacterial populations"

- In the figure legend for Fig. 2, can it be made more clear that the FIT tests are dashed lines

- Could color samples be used for Fig. 1B? It is difficult to tell the different categories in the blob of FIT = 0 data on the left hand side of the figure.

- The test cutoffs are not clearly marked on Figure 1.

- Line 182 appears to be a typo - should be fig S3;

- Line 188 again should be fig S3, not S4

- Legend for Figure S1, last line, should be 'adenomas and normal samples'.

- Please could you also indicate within your methods section whether your research conformed to the Helsinki Declaration

- Please also add a separate section for ‘Competing interests’ according to our instructions for authors