The primary concerns of the three reviewers were that the study lacked in impact and relevance. The reviewers pointed out areas in which our analysis and rationale for pooling samples needed more clarity and explanation. Finally, the reviewers suggested alternative approaches for OTU assignment and Random Forest model training. With that in mind we have added details to the methods and discussion to address these concerns.

Based upon suggestions from the reviewers we have made the following specific changes:

Reviewer #1 (Reviewer Comments to the Author):

The study "Spatial variation of the native colon microbiota in healthy adults" by Kaitlin J. Flynn and colleagues investigates the microbial population of the proximal and distal colon mucosa and luminal contents using a 16S rRNA approach. The study is well designed for the question being asked by the investigators and the results are well described. While the study is technically well done, it does not address an important gap and does not provide a significant advance. The primary difference between the current study and previous studies (Am J. Physiol Gastrointest Liver Physiol 2010 Dec; 299(6): G1266-G1275 Nature Reviews Microbiology 14, 20-32 2016), which have described both temporal and radial differences in community structure, is the collection of samples in the absence of colonoscopy preparation. This is indeed important as colonoscopy preparation can affect microbial community. I commend the authors for this approach, but the relevance of the findings is still unclear and how the differences in findings from the current study improve our understanding of the role of specific microbes as compared to previous studies. The singular descriptive finding in the study lacks novelty and appears premature for publication. However, a comparison with samples collected similarly from a disease cohort would help advance the field as it may more accurately reflect microbiota changes and may shed light on previously missed associations.

The authors only do 16S-based compositional analysis, which as expected shows greater inter-individual variation that the temporal changes within an individual, hence looking at OTUs between the two sites from the collection of subjects would be less relevant than individual-specific differences in sites. The ability to classify samples using machine learning though a good bioinformatic exercise does not add to our knowledge as we are already aware of the site from which the samples were collected. It would be more informative to know if functional characteristics of the microbiome, such as with metagenomics or metatranscriptomics, are different across the length of the colon.

We thank the reviewer for their comments regarding the limited size and scope of this study. The reviewer questions the relevance of studying samples collected from an unprepared colonoscopy while lauding our efforts to collect these samples. The immediate relevance is that we were able to match luminal samples from mucosal samples for the same patients. While it is true that a number of other studies have characterized the mucosal biogeography, those studies are generally small and do not have the ability to characterize the lumen. In fact, the paper the reviewer cites from Hu et al. (doi: 10.1152/ajpgi.00357.2010) sampled four individuals with prepped colonoscopies. Earlier work from Eckburg et al sampled three individuals with prepped colonoscopies (doi: 10.1126/science.1110591). Beyond the ability to characterize matched luminal and mucosal samples, we were able to use more modern techniques to deeply sample each sample. The Hu et al. study used TRFLP, which provides minimal taxonomic information and the Eckburg et al. study sampled a total of 13,355 sequences from a total of seven sites and three subjects. In contrast, our analysis used 423,100 sequences from a total of five sites and 20 subjects.

There is a dearth of disease-cohort samples collected using unprepared colonoscopy in the literature. This is primarily because it can be difficult to get disease-cohort patients to consent to a procedure unnecessary to their treatment, as well as to get IRB approval for unprepared colonoscopy on such patients. Indeed, the only disease-cohort unprepared colonoscopy study used only flexible sigmoidoscopy to sample distal disease sites and did not access deeper regions of the colon (Rangel et al 2015). Of the few published studies, their collection and sample processing steps differ from ours in a way that rigorous direct comparison of that dataset to ours would not be possible. Obtaining the IRB permissions for this study of healthy individuals required more than a year; we have been told that doing a similar study on a diseased population would have been impossible. Thus we find value in our healthy subject sample set not in “improving our understanding of the role of specific microbes” but rather in profiling of the microbes at each site in the (unprepped) colon.

We agree that alternative ‘omics approaches would be valuable in future studies. However, these are limited by scientific and methodological difficulties. Metagenomic shotgun sequencing could be feasible, but this would yield a list of genes, half of which with no known function and we would not know which genes were relevant at the site in question. Although metatranscriptomics could shed light on the genes that are expressed, this approach would be only be possible from the luminal and fecal samples; it was not possible to biopsy sufficient biomass to extract sufficient RNA. We have added a note about future studies to the end of the discussion on line 340.

The authors report they found certain disease-associated taxa in their analysis which would be expected as the majority of the studies reporting microbiota associations report changes in relative abundance of these bacteria between controls and disease suggesting a quantitative difference; hence just the presence of known disease associated bacteria is not surprising. The authors dedicate two paragraphs reviewing the role of Fusobacterium, which was seen in their samples but the current study does not improve our understanding of the role of Fusobacterium in diseases of the colon. Overall, the discussion is too long and often tangential.

We emphasized the finding of disease-associated taxa in healthy individuals in context with previous reports of their prevalence not to suggest potential roles of these bacteria in disease but to highlight how our results showing a location preference for these bacteria in a healthy cohort may precede the development of these diseases. We reduced the amount of discussion spent on Fusobacterium species beginning on line 297.

Reviewer #2 (Reviewer Comments to the Author):

Nice and important study, but the role/presence of biofilms is neglected. Maybe in view of their importance in microbiome-mediated cancer progression (e.g., Li et al. Bacterial Biofilms in Colorectal Cancer Initiation and Progression. Trends Mol Med. 2017 Jan;23(1):18-30), the author should discuss this a bit further and indicate the need for further studies in healthy subjects addressing their absence or presence.

We thank the reviewer for their compliments on our study. We have expanded our discussion on microbial biofilms beginning on line 314.

Reviewer #3 (Reviewer Comments to the Author):

Flynn et al. characterized the phylogenetic (16S rRNA gene amplicon-based) composition of the microbiota of the proximal and distal colonic mucosa and lumen in 20 healthy subjects in order to understand better the spatial variation of the bacterial communities, motivated by an interest in establishing reference data for comparison with those from disease states affecting the colon. Microbial community biogeography in humans is an important topic for which there is a dearth of good data, especially from studies designed with careful consideration of sampling technique, spatial scale and local physiology. Others have already contributed useful data and findings over the past few decades. This study by Flynn et al. has a number of attractive features, including the use of 'unprepped' subjects (minimal prior disturbance of the gut) (and a reasonable number of subjects at that), simultaneous sampling of mucosa and lumen, and a near-contemporaneous stool sample. Among the interesting findings, the authors provide evidence of community distinctness between proximal and distal colonic mucosa, but not lumens, and greater similarity of stool to distal lumen than of stool to any other sampled site. At the same time, there are some weaknesses to this study, including limited sampling along the longitudinal axis of the colon (only two sites), which leaves unanswered some important questions about spatial patterning and underlying explanatory factors, a week delay between stool collection and other sample collection, and suboptimal choices about data analysis.

1. Given the difference in community structure at the proximal and distal colonic mucosa as reported here, it is disappointing that additional sites were sampled along the longitudinal axis of the colon! The design of the study and the introduction seem to assume that the only geographic issue of interest in the colon is proximal versus distal. But there are both theoretical and anatomic/physiological reasons to postulate that the underlying microbial biogeography is more spatially nuanced and interesting. What about patchiness, as some forms of IBD might predict? What about continuous or discontinuous longitudinal gradients, as the ecology of directional flow systems might suggest? What about circumferential biogeographic patterns, as might be expected from the anatomy of blood supply and lymphatics? What about the cecum, which is the preferred site of some colonic disease? Given that the really tough logistical hurdles had already been overcome (getting healthy consented subjects into the endoscopy suite), it is too bad that a greater degree of spatial/mucosal sampling wasn't undertaken.

We agree with the reviewer that a more thorough sampling approach from subjects undergoing unprepared colonoscopy would be interesting. As we mentioned above, obtaining IRB permission to obtain paired luminal and mucosal samples from subjects undergoing an unprepped colonoscopy was not trivial and increasing the number of samples obtained along the colon would have only made this more difficult. We have highlighted these approaches as possible future directions on line 341.

2. A 2005 study of microbial biogeography at 6 different mucosal sites along the length of the colon in each of 3 healthy subjects (ref 21 in Flynn et al) yielded data that suggested

the possibility of patchiness. The authors might comment upon these findings in this prior study and whether their own study design was suitably organized and powered to have detected a pattern of within-subject patchiness. Here, 'patchiness' should be formally defined and addressed according to biogeographical theory and method.

We have expanded our discussion of spatial ‘patchiness’ in the discussion beginning on line 264.

3. Where exactly in the proximal colon were mucosal and luminal samples obtained? The authors are reasonably precise about the distal location (25 cm of flexible sigmoidoscope), but give no information about how the proximal site was chosen in each subject, and how much variation there might have been in these proximal locations. How far along the pediatric colonoscope??

We have added information about how the proximal site was chosen in each subject and the variation (in cm) in the locations in the methods section beginning on line 133.

4. The 20 subjects were said to have had no antibiotic use during the 3 months prior to sample collection. The difficulty of finding subjects with greater periods of abstinence is understood, but because some antibiotics can have effects that last far longer than 3 months, it would be important to provide information on antibiotic use in the last year for each subject, as well as medical history focused on conditions that typically prompt frequent or extended antibiotic use.

For study recruitment, no antibiotic use within 3 months prior to participation was a criterion. However in practice, 18/20 study subjects had not used antibiotics within the last year, and the remaining 2 had not in the past 6 months. This has been clarified in the methods section beginning on line 116.

5. How were the amounts and consistency of fecal material standardized from sample to sample (especially since consistency is a known important source of variation)?

Changes in consistency were a criterion for proximal sampling and effort was made to take only formed stool from the distal colon and unformed stool from the proximal colon. We standardized the amount of DNA used for input of 16S rRNA sequencing after DNA isolation from the samples to control for differences in samples and have emphasized this in the manuscript methods beginning on line 140.

6. Why were the stool samples collected one week prior to the endoscopic procedures, and not close in time? Stool microbiota structure can vary over the course of a week. How do the authors suggest that this variation be compared to other degrees of within-subject sample variation?

Stool samples were collected *up to* one week prior to the endoscopic procedures and most were collected within a day or two prior to procedure. This was necessary because frequency of stool evacuation is variable among healthy subjects and thus we could not rely on a natural stool evacuation event on the procedure day. Furthermore, although the gut microbiota is dynamic, nearly all studies that have addressed intra-personal temporal variation have found that the amount of change to be expected over a day is quite small. We have clarified the sampling procedure in the methods as well beginning on line 120.

7. Page 6--The identification of OTUs using a % sequence similarity threshold/cutoff is now recognized as importantly flawed (it fails to exclude reads with sequencing errors, and excludes important 'real' reads). This clustering-by-fixed-cutoff approach (used by the authors) is clearly out-performed by what is now a preferred method: 'Dada2' creates a model of sequencing errors and error rates from the raw sequencing data and then tests each read against the null hypothesis for this model, enabling statistical inference of real sequences (Nature Methods 13:581-3, 2016). The much greater 'resolving power' of this method gives more reliable results (especially given the interest of the authors in identifying taxa) and can reveal underlying biology and ecology.

Unfortunately, we must disagree with the reviewer that using a percent sequence similarity threshold is “importantly flawed”. We would contend that the rush to using single nucleotide variants to define an OTU is flawed and introduces a number of important risks that are underappreciated by those pushing the method. Ultimately, the goal of pushing the field to a high similarity threshold is an attempt to get 16S rRNA gene sequences to do something it is not capable of doing. Specifically, 16S rRNA gene fragment sequences cannot delineate bacterial species and cannot tell us about phenotype. If scientists have these types of questions there are far more powerful tools at their disposal than debating the appropriate threshold for defining OTUs such as cultivation, phenotypic testing, and genomic sequencing. There are several risks using data from 16S rRNA gene sequences in this manner. First, there is a real risk of splitting the 16S rRNA gene copies from a single genome into multiple bins. As an example, *E. coli* ATCC 70096 has 7 copies of the 16S rRNA gene and 6 of these are different from each other in the full-length version of the gene. Fortunately, within the V4 region the 7 copies are identical. Alternatively, *Staphylococcus aureus* ATCC BAA-1718 and *Staphylococcus epidermidis* ATCC 12228 both have 5 copies of the 16S rRNA gene. Considering the V4 region of these species, 4 of the 5 copies in each genome are identical between the two species. The remaining S. aureus copy is 1 nt different from the other *S aureus* copies; however, the remaining *S. epidermidis* copy is 1.7 and 2.0% different from the other *S. epidermidis* and *S. aureus* copies. The less restrictive threshold would lump the two species together; however, the more restrictive threshold suggested by the oligotype proponents would generate 3 OTUs. None of these reflect the biology they claim and the method would split sequences from the same strain into different OTUs. These types of examples abound in gut-associated bacteria who generally have more than 5 copies of the 16S rRNA gene and they are never identical. Second, there is a real risk that artifacts of sequencing noise could be binned into their own OTU. Although tools like Dada2 and mothur do an exceptional job of mitigating this risk, PCR and sequencing errors are not random and it is possible for these artifacts to accumulate a number of reads leading to the over interpretation of the biological significance of such an OTU. Finally, there is a risk with oligotype-based approaches that one will artificially inflate the differences between communities. Given the levels of patchiness seen in human microbiome data, in general, and in our data, in particular, we contend that the OTU definition we have selected is appropriate for our set of questions. As our analysis reflects, we believe that it would make sense to take a more guarded recommendation than to make dogmatic pronouncements about using high thresholds.

8. Pages 6-7--Rarefaction is problematic, and some would judge to be statistically 'inadmissible' since it requires omission of valid data and undermines the performance

of downstream methods (PLoS Comput Biol 10: e1003531, 2014). A preferred approach for dealing with libraries of different sizes is described in this citation, and involves a variance stabilization technique that has been validated and used historically to address this problem with other similar types of data, e.g., RNA-Seq data.

It should be highlighted that the approach described by McMurdie et al. is not a universally accepted approach for microbiome or ecological analyses. Rarefaction is widely used throughout ecology to control for uneven sampling effort in alpha and beta diversity analyses. With microbiome data, even the best of pipelines generates spurious OTUs and these OTUs increase with sampling effort. Therefore, rarefaction controls for both sampling effort and sequencing artifacts. The analogy to RNA-Seq data is close, but misses a key difference. In RNA-Seq data, each sample is drawn from the same genetic population. In other words, RNA-Seq done on *E. coli* will be based on the same genes. In data such as ours, the taxonomic membership of each sample varies considerably. Thus, a zero count in a RNA-Seq study may be interpreted to mean that the expression was below the limit of detection. In our case, a zero count may be interpreted the same way or as meaning that the population was not present. There is a risk of falsely identifying significant OTUs if some samples (e.g. luminal) generate more sequence reads than others (e.g. luminal) – the rare OTUs in the luminal samples may be identified as discriminating between sites not because their relative abundance is different, but because more sequences were sampled. It is worth noting that the benchmark dataset used by McMurdie et al was not significantly patchy and the samples were all relatively consistent in the number of reads per sample. Finally, the DeSeq approach suggested by McMurdie et al. is effective for identifying individual OTUs that discriminate between groups. Our use of similar methods identified very few OTUs that had a significant difference between groups. This is why we have used the Random Forest approach, which takes into account multiple OTUs at a time when making a classification. Regardless of one’s opinion of the McMurdie approach, methods of handling the zeros are not readily available for building Random Forest models. By our approach, the risk is that we aren’t identifying enough populations, not that we have identified too many.

9. Were separate lumen-mucosa and proximal-distal sample comparisons undertaken when controlled for subject, as well as undertaken across subjects? If not, they should be. Given the dominant contribution of individual to inter-sample variation, it would be important then to subtract this source (control for individual), when considering other sources of variation.

As described in the results, we did examine individual comparisons of sites within a subject but no clear pattern emerged. In the diversity analyses (Figure 3A-B), the diversity is only comparing between sites of the same patient, and thus the overall distribution shows the range of differences in diversity among the 20 subjects. This is how we have chosen to represent the variation within subjects as well as in Figure 3C. The Random Forest models compared the locations across all subjects (Figure 4).

10. Page 7--Validation of the findings from the Random Forest model was performed with a leave-one-out approach. This is generally sub-optimal. More robust validation is achieved with the use of separate learning and testing data sets. Ideally, these two data sets are derived from different populations of subjects.

Given the small n of our dataset and how samples were acquired, we are unable to split the data in to completely separate learning and testing datasets and still generate a meaningful model, hence why we have employed a leave-one-out-approach. We have additionally used 10-fold cross validation to further estimate the prediction error of the model. 10-fold cross validation partitions the dataset into 10 different testing and training portions and running the model 10 times and averages the result. This is to prevent one test or training set from biasing the resultant model. By combining the leave-one-out approach and 10-fold cross validation we address overfitting problems despite our small dataset. We have clarified our approach in the methods beginning on line 152.

11. Can the authors provide a citation for the statement that proximal mucosa contains the highest oxygen concentrations of the colon (lines 268-9)? Does the anatomy of the colonic blood supply support this statement?

This sentence has been reworded to reflect that the highest oxygen concentrations of the colon are in the proximal colon and thus facultative anaerobes found in the mucosa are consistent with an oxygenated microenvironment line 281.

12. Figure 1 is unnecessary, but if included, it certainly doesn't need to be a main figure.

We respectfully disagree with the reviewer as the approach of unprepared colonoscopy is not standard and thus the photos obtained from the sampling locations are informative.

13. There are a number of typos and grammatical errors. For example, in Intro: line 46, missing "in"; line 50, "tumors" can be deleted (since CRC is the subject); line 86, to what does "latter" refer? Also, lines 282-3, "more" seems to be missing in front of "variable".

The minor grammatical and punctuation errors have been corrected throughout the manuscript.