Reviewer #1 (Comments for the Author):

Major Comments P.5 (line 81): The bioinformatics behind the statistical analyses are important. Include a brief summary in the main text. Specifically, it is important to clarify that you focused on studies that collected targeted 16S sequence data (not shotgun metagenomes). Also briefly mention that you did OTU clustering from scratch (per study and pooled?). Here or in the methods, clarify how alpha and beta diversity statistics were estimated.

We have added text to address these comments (L81 and L91).

P.7 (lines 134-136): To overcome challenges with non-overlapping 16S regions, might it be possible to predict taxonomy of OTUs (which it appears you have already done for phylum level analyses) and then do a pooled community structure analysis on the taxa rather than OTUs? Alternatively, perhaps tools designed for identifying OTUs from non-overlapping shotgun marker gene reads might work.

We did this in the random forest analysis at the phylum, class, order, family and genus levels and did not see a difference in the results relative to the OTU-based analysis. We have added text at L161 to L165

P.8 (line 162): What variability was used in the calculations (presumably that observed in the studies)? For designing future studies, it would be great to investigate the sensitivity of the results to the amount of between-sample variance, since variance does differ between studies (e.g., is higher in rural populations and is lower between moms and their newborns than between unrelated individuals).

We used the observed variation in the data from each study. Added a comment in discussion to address cohorts that would result in different levels of interpersonal variation.

P.10 (line 204): When discussing biological relevance, it would be useful to also mention that relative abundance may not be the relevant factor for the host. Imagine an OTU (A) that produces a metabolite with a physiological effect on obesity (or whatever host trait one is studying). Suppose the number of cells and the amount of metabolite produced remains constant, but the number of cells for some other OTU (B) goes up by an order of magnitude. The relative abundance of A in the community has gone down, but this may be irrelevant to the molecular/cellular mechanism affecting obesity, especially if B has no relationship to A or obesity (e.g., it occupies a distinct niche in the gut). The same argument probably holds for alpha diversity metrics based on relative abundances - the host likely cares about specific components of the diversity not diversity per se.

We agree with this sentiment and included text at L258-267 to point out that there are clear examples where changing the microbiome changes BMI profile. We have expanded on this point as suggested by the reviewer below.

P.12 (line 244): Where can the pipeline be obtained? This info is in the methods but should probably be highlighted.

We have added a link to the GitHub repository that contains the pipeline for the analysis and this manuscript.

P.12 (lines 265-267): Elaborate further on this point. The right biomarker may be a protein or pathway. Even if this is taxonomically restricted, it may be hard to detect with community-level statistics or even OTU relative abundances. In addition to the issue noted above regarding relative abundance, species-level OTUs can still miss functional variability due to large differences in gene content among strains with identical or nearly identical 16S sequences.

We have added text to this paragraph to expand upon this thought.

Minor Comments P.4 (lines 53-56): repeated text and typos: "mechanistic studies using animal models that were manipulated with antibiotics or colonization with varied communities were manipulated with antibiotics or underwent colonization with varied communities appears to support the association since these manipulation yielded"

This has been fixed.

P.6 (line 100) and elsewhere: When first referring to a study by name (e.g., Walters, Finucane, Goodrich) provide the citation. This info is in Table 1, but it would be helpful to provide in the text as well.

We have added an explicit reference to the first time we mention the Walters and Finucane studies at L50. Also, to respond to a comment from Reviewer 2, we have added the reference numbers to Table 1. This should make the connection between the studies and specific references more clear.

P.6 (line 103): In the random effects model does study have a random intercept or slope (or both)?

Each study has a random intercept and slope.

P.6 (line 88): What is the overlap of studies with the Walters and Finucane analyses (could move or copy this from the discussion)?

We have added text at L89 to L91 to indicate which of our studies overlapped with those from the Walters and Finucane.

P.6 (line 88): What is the range of sample size per study (in Table 1, but would be helpful to give this overview statistic in the text)?

Given the broad variation in the number of subjects per study, it does not seem appropriate to reduce that variation to a single parameter or a range of values. The table lists the sample sizes for the ten studies.

P.7 (lines 122-123): Clarify if ranges on relative risk are 95% confidence intervals or some other statistic.

Added “95% CI” to each set of parentheses to indicate that these are the 95% confidence interval.

P.7 (line 123): Fix plural: “were no significant difference”

This has been fixed.

P.7 (lines 132-133): Explain why directionality cannot be inferred.

We have added text to clarify the point. Basically, in a univariate analysis, we can say that a group of people are taller or shorter than another, but we cannot do the analogous comparison to community structures since their relationship is in an no-dimensional space.

P.13 (line 281): Is this algorithm the same as agglomerative hierarchical clustering with average linkage? Also, provide some more detail on how the read counts for OTUs were used to estimate the various alpha and beta diversity statistics.

Yes it is. There are multiple names for the same algorithm. A citation is provided to make it clear what we are referring to. The number of reads that clustered within each OTU or taxonomic group were used to calculate the various alpha and beta-diversity metrics. We have added a citation to Magurran (2003) that describes the forumlas that were used to calculate the various parameters.

P.14 (line 297): Typo: "each study within a study"

This has been fixed.

Reviewer #2 (Comments for the Author):

General comments: Overall this paper is a useful meta analysis of obesity and the microbiome. I have some miscellaneous comments that are below. In addition I have one more important concern that I believe needs to be addressed.

More important concern: Given that their analysis depends on them compiling data from multiple studies and reanalyzing that data, I believe it is necessary for these authors to make sure that all components of this analysis are reproducible. I commend the authors for their commitment to sharing code for all the work. However, I have a few concerns about the data. First, some of the data is from a personal web site. I would recommend that the authors here either repost that data to a more sustainable site (e.g., Figshare), get the original producers of the data to post it to a more sustainable site, or come up with some other solution. Again, I commend the authors for their efforts in openness and reproducibility, I am just concerned that some aspects of this paper will not be reproducible.

We appreciate the reviewer's concern regarding the openness of the underlying data used in this study. As far as we can tell all of the sequence data is available either through the ENA, SRA, or dbGaP except for the Turnbaugh and Schubert datasets. Given the age of this dataset and the lack of access to the original sff files (we've asked previously) it is unlikely that continued requests to the Gordon lab will get these data to be made any more public. Furthermore, this is a keystone study in this research area because it was the original study to claim a relationship. The Schubert dataset is now available on the SRA (accession SRP078489). The reviewer is correct that some of the metadata is not publicly available (i.e. Wu and Zeevi). These were obtained from the original authors. We have encouraged the authors to make these data more available.

Regardless of the outcome of that encouragement, we have posted the metadata for each study in the GitHub repository for this study. This is obviously a complex problem and we can only do our best to encourage our colleagues and follow best practices ourselves. Ultimately, we do not think it is necessary to censor datasets from the study because their data are not publicly available.

Additional concerns and comments: L47 "individuals had a lower diversity than lean individuals (6)." Specify what kind of diversity

We have specified that it was alpha-diversity. It is actually unclear what type of diversity the original authors intended. Given that richness and diversity are often used interchangeably, it is difficult to know with certainty what was meant.

L54 "with antibiotics or colonization with varied communities were manipulated with antibiotics" Typo? Should this be "that were manipulated?"

We corrected this sentence.

L73 "Literature Review and Study Inclusion. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to identify studies to include in our meta-analysis (14)." Would be good to say something about whether this is considered an important approach in meta-analysis

We have edited the sentence to make it clear that following the PRISMA guidelines is necessary to prevent inclusion bias and to be considered a robust study.

L77 "we searched PubMed for original research studies that involved studying obesity and the..." Given the importance of covering the literature well in meta-analyses it would be good to say how Pubmed was searched.

L78 "We identified ten additional studies" How?

The exact approach we used has been clarified in the Supplemental Materials.

L80 "and obesity. We then manually curated the 196 studies to select." Numbers don't add up - 187 + 10 = 197. I looked through the Figure and I am not sure this explains all the numbers either.

We corrected this error in the new review.

L102 "other studies appeared to have the same trend, albeit the differences were not statistically significant." Quibble about wording here. If the result is not significant - is it a trend?

We appreciate the reviewers' concern, but alternative wording (e.g. effect sizes had the same sign) was clunky and effectively said the same as "trend". We do not think that a trend has to be significant to be a trend. We are also critical of others that claim pseudo-significance from things that are "trending", but feel we are being cautious here since we then do the aggregate test to show that the trending is in fact significant when considered across all of the studies.

L110 "significant lower diversity than non-obese individuals; however, it is questionable whether the difference is biologically significant." Some justification for this statement is needed.

L126 "obese, it is questionable whether that risk is biologically or clinically relevant." Again, Some justification for this statement is needed.

We expand upon these statements at L204 where we discuss the need to develop guidelines as to what is a biologically significant difference. There are likely some researchers who would contend that any statistically significant difference in diversity is also biologically significant; however, it is also possible for a study to be overpowered so

that it could yield statistically significant differences that are not biologically significant. Our discussion at L204 comments on this distinction and the need to develop better theory around what is biologically significant for microbiome studies.

L164 "each of the studies (Figures 6, S3-S8). Although there is no biological rationale for these effect sizes, they represent a range that is plausible." It would help to clarify what is meant here by "range that is plausible"?

We have clarified that by plausible, we mean that the range includes values that most microbiome researchers consider to be biologically meaningful - even if they do not know what it means. Furthermore, most of the studies were sufficiently powered to detect effect sizes of 15%.

L193 "datasets. This analysis demonstrated that the ability to reliably classify individuals as being obese based on the composition of their microbiome was limited" I think it would be good here to clarify that this is really the ability to classify based solely on microbiome data. It might be possible to do this much better if individuals were subclassified (e.g., by genotype, age, gender, etc)

We have highlighted the text to indicate that the analysis was based solely on microbiome data.

Some references could / should be added: lme4 (v.1.1-12) R package, vegan (v.2.3-5) R package, AUCRF (1.1) R package, pROC (1.8) R package, pwr (1.1-3) R package

We have added references for lme4, pROC, and AUCRF. The others have not been published.

Table 1 - would be good to add the references for the studies

We have added the citation number for each study to Table 1.

Figure 1 - some of the boxes in the flow chart are cutoff.

We have added whitespace around the boxes on the right side of the figure.