THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

March 19, 2019

The following is our point-by-point response to reviewer concerns.

Reviewer 1

The article by Shaiber and Eren is an important wake up call to the growing problem of incompletely curated genomes assembled from metagenomes. The authors pinpoint the scientific problems with such datasets, that can lead to incorrect conclusions and propagate in further studies and in the literature. Unfortunately, most such issues likely remain undiscovered and, even when identified, public databases do not offer effective mechanisms for correction.

The field of metagenomics-based genome reconstruction continues to evolve, and there are a plenthora of approaches and software to handle the data, from cleaning up the sequence reads, assembly, to binning and iterative genome curation. Research groups that generate metagenomic data but are not experts with the details and limitations of the various bioinformatics code used to process the data could find themselves selecting ineffective strategies, that may work well for some microbes but not for others. While a standardized "pipeline" will not be immediately adopted by the entire metagenomics community, groups that are at the forefront of developing metagenomic data analysis software (such as Eren's), should actively educate the users on the state of the field and promote effective approaches for metagenomic data analysis, from start to finish. Specifically, because this manuscript is directly addressing the problem of ineffective genome curation, the authors have the opportunity (and responsibility) to do so and, for the most part, they do.

We thank the reviewer for sharing this heavy burden with us by sharing their insights to improve this letter.

However, while the code and commands to analyze the specific data from the Espinoza study are provided, would those be directly applicable to other datasets, not only to re-analyze genomes but also to generate them?

From multiple assembly strategies to methodologies for mapping and binning, there are many ways to generate genomes from metagenomes. The method followed by Espinoza et al., the co-assembly of metagenomes to generate contigs and the use of a binning algorithm to reconstruct genome bins, is one of the common ones, and we agree with the general approach. Our disagreement stems from the fact that the authors did not make any attempt to refine resulting genomes despite their composite nature, and they proceeded to submit them to a public genome repository. Hence, our attempt was to remind the community that this practice will have an influence on the quality of public databases that support science and demonstrate how the refinement step could be improved to avoid similar outcomes. But we do agree with the point our reviewer makes about the fact that our online workflow may be too specific to the refinement of Espinoza et al. genomes. Our revision includes a much more comprehensive and easier-to-follow tutorial that will guide others to be able to refine their own MAGs in a similar fashion. Our updated online material (http://merenlab.org/data/refining-espinoza-mags/) has a longer introduction, contains a step-by-step workflow with improved details and tips for those who wish to scrutinize their genomic discoveries. We thank the reviewer for their suggestion.

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

In the opinion of the authors, were there specific incorrect/ineffective steps used in the Espinoza study, and what would be the better alternatives? I realize there is a space constraint but that could be accomplished online.

In the online analysis flow it appears several other genome bins were analyzed but not discussed in this paper. Were they also found to be contaminated or was that analysis not completed?

The reviewer is correct. Our analyses were limited to a small subset of Espinoza et al. genomes. The reason behind this was our desire to use our limited space strategically. Most of the heavily contaminated genomes in Espinoza et al. study have very high-quality representatives in comprehensive resources such as the Human Oral Microbiome Database, hence other researchers will unlikely rely on Espinoza et al genomes. In contrast, the three MAGs we focused on in particular are those that belong to the poorly understood branches of the human oral microbiome and will likely be used by others in phylogenomic, pangenomic, and/or metagenomic read recruitment analyses for comparative purposes. We now have clarified this in our online material and removed the *Alloprevotella* genomes from our analyses to avoid any confusion.

Essentially, would all the genomes reported by Espinoza et all likely need to be reanalyzed before used by others? I realize that is speculative, but if certain steps of their analysis were ineffective, that may be needed.

While we agree with the reviewer that a more complete reanalysis of each genome reported by Espinoza et al. may be necessary, we believe it would be more collegial if our letter did not make any suggestions on this point. We hope that while contributing refined versions of critical genomes in Espinoza et al. study, our letter will raise awareness among microbiologists and those who review findings that emerge from genome-resolved metagenomics regarding the importance of at the very least enforcing the existing quality measurements for cleaner public genome databases.

The bacteria discussed in this manuscript need to be named appropriately throughout the manuscript. Neither TM7, GN02 or SR1 are validly described taxa, and the recent renaming them as Saccharibacteria, Gracilibacteria, Absconditabacteria does not make them valid either. Therefore, the authors should use TM7/ Saccharibacteria, etc, both in text and in the figure.

We agree with the reviewer that the naming is confusing. Furthermore, we believe the purpose of these names is to implement a common language that is supported by molecular and phylogenetic analyses, and the validity of these names are solely defined by the community who use them. We had initially tried to follow the nomenclature that was used by Espinoza et al's publication. However, to address the point made by our reviewer and to help readers to link old names to newly proposed ones, we mention them together in our revised manuscript. We hope the reviewer agrees with our resolution.

We are thankful for the time the reviewer invested in our work.

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

Reviewer 2

This is a conceptually important letter for the community but there are some aspects that I have questions and / or concerns about. These are detailed below.

We thank the reviewer for their time and insightful remarks.

1. Title: Poorly resolved metagenome-assembled genomes contaminate public genome repositories
   The wording here does not precisely reflect what is in the letter. First, I am not sure what is meant by "poorly resolved" here. In the text of the letter use refer to MAGs that "resolve to poorly understood clades" but I assume that is not what they are referring to here but rather to something to do with the quality of the assignment of MAGs to taxonomic groups and poorly assembled MAGs? It would be good to clarify. . Second, the choice of the term contaminate is not ideal. The letter discusses the use of the term "contamination" in the context of assessing MAGs and this is different from the use in the title.

We agree with the reviewer that the wording of the title made it open to misunderstandings. Accordingly, we changed the title to:

> *"Composite metagenome-assembled genomes reduce the quality of public genome repositories"*

We believe this revised title appropriately states our message. When genomes of composite nature (i.e., those that include large portions of sequences that originate from distinct microbial populations) are deposited to public databases as single genomes, they reduce the quality of such repositories as the scientific community no longer knows which genomes to trust and use for future analyses. We agree that our initial submission failed to convey this message clearly. Our revised letter includes substantial changes and new analyses to streamline this point.

2. L12 "88 individuals" would be good to say "human"

We first added the word "human" to clarify, however, "human individuals" sounded redundant. Indeed, the definitions of 'individual' on Merriam-Webster suggest the common use of this term as referral to humans. We elected to keep "individuals" in our text as is especially since the next sentence mentions "human oral cavity". We hope the reviewer agrees with our reasoning.

3. L18. "composite genomes" it would be good to define this here.

In our revised manuscript we clarified the meaning of this term (please see below).

4. L18 "which renders them unsuitable for future comparative genomics studies." I do not think this conclusion is supported. They should be used with caution certainly and should be labelled as such but they still could be used for some comparative studies if used in the right way. The original human reference genome was a composite and was used inappropriately by some but perfectly appropriately by others. I think it is a major overstatement to say these are "unsuitable for future comparative genomics studies."

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

We appreciate the reviewer's suggestion and agree that such remarks should not be made without explicit support in case the implications are not clear to all readers immediately. First, we have replaced the sentence the reviewer mentioned with another one that is clearer, and supported our statement by citing a recent study in which incorrect claims resulting from a composite genome were revealed by refining the genome:

> *"Composite genomes that aggregate sequences originating from multiple distinct populations can yield misleading insights when treated and reported as single genomes (4)."*

To demonstrate how ecological interpretations may be influenced by composite genomes we gave an example from Espinoza et al. study (L35):

> *"For instance, the original MAG III.A recruited a total of 1,849,593 reads from Espinoza et al. metagenomes, however, the most abundant refined III.A genome (MAG III.A.2, Figure 1C), recruited only 629,291 reads."*

This change in the number of mapped reads demonstrates a 3-fold overestimation of the abundance of MAG III.A.

To further demonstrate the consequences of using composite genomes, we performed pangenomic analyses using unrefined and refined version of Espinoza et al MAGs. Our revised Figure 1E now includes the results of this analysis. Our pangenomic analyses with refined MAGs increased the number of single copy core genes by a factor of 200%-1,600%. We added the following section to our revised letter (L29):

> *"A pangenomic analysis of the original and refined MAG III.A genomes with other publicly available Saccharibacteria genomes showed 7-fold increase in the number of single-copy core genes (Figure 1E). These findings demonstrate the potential implications of composite MAGs in comparative genomics studies where single-copy core genes are commonly used to infer diversity, phylogeny, and taxonomy (6)."*

We also analyzed the read recruitment results of MAG IV.B (Figure 1B) to demonstrate how composite genomes influence ecological insights through incorrect prevalence estimates. We measured the 'occurrence' of each refined IV.B MAG across individuals included in the study. To assume a MAG is detected in a sample, we required as low as 25% of its entire content to be covered by at least one short read in mapping results. This analysis showed that refined MAGs co-occurred only in about 10% of the metagenomes, demonstrating their distinct ecology. This striking result which is summarized in the table below did not change when we used more conservative detection thresholds of 50% and 75%. Since the difference in occurrence is visually evident from the figure, we chose to exclude this analysis from the text due to space limitations. We hope these findings help address the reviewer's concerns.

	Num samples with detection		
Detection threshold	MAG IV.B.1	MAG IV.B.2	Num samples refined MAGs co-occur
75%	6	5	1 (10%)
50%	17	13	2 (7%)
25%	32	18	5 (11%)

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

Besides having distinct distributions across the 88 metagenomes of Espinoza et al, the new phylogenomic analysis (Figure 1D) shows that both MAGs we refined from MAG IV.B are closely related to genomes published previously (the figure includes their accession numbers). Analysis of the average nucleotide identity (ANI) across genomes further supported this finding. ANI results (which we also didn't include in the letter) showed that the two MAGs we refined from MAG IV.B are quite distant to each other: only 25% of their genomic content aligned with approximately 80% identity. In contrast, more than 80% of each of the refined MAGs align to the previously published genomes with an identity over 98%.

ANI is emerging as a standard approach to define operational boundaries for microbial 'species', and recent studies suggest that organisms that belong to the same species typically show ≥95% ANI among themselves (see Jain et al. for a recent validation based on available genomes). In the light of these suggestions, our ANI results suggest that populations that were mixed in MAG IV.B originate from two distinct species, if not genera.

5. L21. "To briefly demonstrate the extent of contamination". Contamination is not mentioned previously. How is it being defined here and where in the previous text is this discussed. Do the authors view composite MAGs as contaminants? Or is it the DNA from other taxa that are not supposed to be there that are the contaminants? I truly am not sure from reading this what they mean here.

We changed this sentence to use the term "composite genomes", which is defined earlier in the text:

> *"To briefly demonstrate their composite nature, we refined some of the key Espinoza et al. MAGs through a previously described approach (5) and the data the authors kindly provided (1)."*

6. L23. Gracilibacteria MAGs -it would be useful to have / use IDs to refer to MAGs being discussed since the taxonomy of the MAGs is an inference and an inference in some cases the authors here believe is wrong. Referring to MAGs by some ID would be therefore better than referring to an inferred taxonomy.

Changed to use IDs instead of taxonomy.

7. L23 Just showing that the distinct subtypes of these MAGs have inferred different relative abundance patterns across samples does not mean they have distinct ecology and I would recommend not using that term and just reporting the result (different patterns of relative abundance across samples). Or the authors should explain more why they think this means they have distinct ecology.

Removed "distinct ecology". The sentence now reads:

> *"We found that MAG IV.A, MAG IV.B, and MAG III.A described multiple discrete populations with distinct distribution patterns across individuals (Figure 1)."*

8. L23 related to the above comment, the authors do not provide sufficient explaination for why they think the results presented in Figure 1 demonstrate that the two MAGs in Fig 1 A and 1B in fact should be viewed as composites. I think I understand the authors logic here, and I think I would likely agree with them, but just showing these figures is not enough. Is

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

the analysis that they did robust? What evidence can they present that the way they analyzed the data is not introducing its own artifacts? Again, I think I am likely to support the authors inferences but not from the incredibly limited information presented here.

We thank the reviewer for their skepticism. We have now cited multiple studies to support our claims, and previous applications of methods we have used that demonstrate their robustness. The efficacy of our approach to remove contamination is supported by remarkable change in single-copy core gene estimated contamination levels, which is a standard measure in the microbiome field (Bowers et al. (2017)). Our visualizations of the distribution patterns across MAGs offer striking visual evidence for these statistics by explicitly displaying the underlying data. In addition, we have now added a citation for Delmont and Eren (2016), in which a detailed description is provided for the method that we use for the refinement of the Espinoza et al. MAGs. Besides, the updated online material includes extensive detail for anyone to reproduce our findings.

9. L24-27. The authors basically state that the taxonomic assignment for the MAG in Figure 1A is wrong and refer to ribosomal protein analysis that "resolved both of the new populations in Figure 1A to the candidate phylum Absconditabacteria (formerly SR1), and not to Gracilibacteria" Yet as far as I can tell no evidence is presented for this statement nor any discussion of how or why this type of analysis should be considered better than what the original authors did. This therefore does not, as the authors here state reveal "the importance of using multiple approaches to verify taxonomic assignments." It might, if they showed their approach was better or even a reasonably alternative. But since they do not present their results in any way to examine, nor discuss why it might be good / better, this statement is just not supported.

We thank the reviewer for pointing this out, and we agree with their criticism. Our revised manuscript now includes the phylogenomic evidence that supports our statement (Figure 1D).

10. L30-32. "Nevertheless, reporting multiple populations with distinct distribution patterns as single genomes should be avoided to prevent misleading ecological and functional insights." As with comment 4 above, I do not believe this statement is supported by any evidence presented. Certainly, composite MAGs can be problemmatical in many ways. However, does this mean they should be avoided per se and thus more evidence for the problems that would come from releasing such MAGs needs to be presented here to justify such a recommendation. Alternatively if the authors want to state an opinion, I think it might be OK if they said "We believer that there are too many problems with releasing such MAGs" but to state it here as though it is the only conclusion one can make, is just wrong.

We hope the reviewer finds the new evidence we provided in our revised letter, online resource, and our response to be supportive of our statement. The importance of ***not*** reporting multiple microbial populations that have distinct distribution patterns across habitats as single genomes for other scientists to use is paramount. There may be situations where composite genomes can offer invaluable insights but reporting them as single genomes should not be tolerated as an acceptable practice under any circumstance, especially when these composite genomes can be cleaned up so easily as our letter demonstrates. The fact that the study by Espinoza et al. was reviewed and published with these inappropriately finalized genomes despite the simple community guidelines and available tools is evidence that we need to urge the microbiologists firmly to scrutinize similar studies further. It is not a matter of belief that composite genomes that are presented as single genomes will lead to incorrect ecological and functional insights. Our revised letter now includes

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

more evidence within the confined space afforded by the article type for this point, additional citations to demonstrate previous examples of mishaps due to composite genomes, and an improved online documentation to clarify the common solutions better. We hope the additional evidence will pursue the reviewer to reconsider, and they will agree that it would not be appropriate to label these clear and previously documented risks to scientific discourse as mere "beliefs" or "opinions".

11. L33. Use of contamination here - need to make sure it is the same definition as elsewhere and that this is clearly defined somewhere.

We added a definition, which is consistent with the one used by Espinoza et al.:

> *"MAGs that suffer from extensive contamination, as measured by the redundancy of single copy core genes, (Figure 1 in this letter and Table 1 in Espinoza et al.)…"*

12. L34-35. "Available as single genomes in public databases of the National Center for Biotechnology Information (NCBI)." This to me seems to be one of the key aspects of the issue with MAGs and could be explanded. It seems that there are in fact real problems if NCBI is labelling composite MAGs as "genomes". That is not a problem with MAGs or composite MAGs or the authors of this paper but rather a problem with NCBI mislabelling things. I would suggest most of a focus on this issue rather than on making claims about how composite MAGs should not be released when it seems that the actual issue may be with NCBI mislabelling and then misuse of that mislabelled data.

Composite genomes that explain multiple population genomes with distinct ecology are improper units of diversity to study microbial life. We agree with the reviewer on NCBI's shortcomings to label poorly finalized genomes, however we believe that is an independent problem. If the community recognizes the former, the significance of the latter problem would be much less urgent.

13. L59-60 "But we also must keep in mind that we are the only stewards of the quality and purity of our public databases." Who exactly is the "we" being referred to here? The authors? The community? It would be good to clarify.

We removed the sentence.

14. L63 "extensive contamination of public resources " It seems that this is a different use of "contamination" than in the other parts of the paper except the title. It would be helpful to not use the term in two different ways but rather to come up with another term to use for one of the uses.

We removed this part of the sentence.

15. L63-66 "will yield long-lasting adverse effects such as misleading insights into the taxonomy and functional potential of novel groups and reduced trust among scientists in findings that emerge from genome-resolved metagenomics.. I think that if this type of statement is to be made it should be presented as an opinion not in essence a fact. This something like "We believe this will lead to long lasting ..." Alternatively more evidence for actual misleading insights from such MAGs would be needed.

THE UNIVERSITY OF CHICAGO MEDICINE

A. Murat Eren, Ph.D.
ASSISTANT PROFESSOR, DEPARTMENT OF MEDICINE

Knapp Center for Biomedical Discovery
900 E. 57th Street, Mailbox 9, Room 9118, Chicago, IL 60637
P: +1-773-702-5935 / F: +1-773-702-2281 / meren@uchicago.edu

Our revised letter now includes evidence for evolutionary and ecological implications of composite MAGs with new analyses and citations.

16. I like that the authors analyzed a preprint. I am however, concerned that the letter refers to the preprint only when a final version apparently based on this preprint is now published. It would be good for the authors to make some statement regarding the final version (e.g., maybe no changes were made in areas that affect this letter). But the final version should not be ignored now that it is out.

Our initial submission mistakenly cited the pre-print rather than the final study even though all our analyses used the data and findings that appeared in the final version of Espinoza et al. study. We now have fixed the citation. We thank the reviewer for their time and attention to this critical detail.