Response to Reviewers

**General Notes**

We thank the reviewers for their comments and we appreciate their positive support. We have made important improvements to our manuscript. Many of the comments we received suggested that we should improve our descriptions to make our goals and methods more clear; we have significantly altered the text to use a more consistent and understandable language throughout.

In the interest of reproducibility we have added or rebuilt several figures and we will deposit the code used to generate these figures in our public GitHub repository. Additionally we have ensured that all of our results include comparisons between the output of our algorithm and standard linked reads which we believe is the most apt comparison for our algorithm in the absence of a similar class.

We have more clearly defined the scope of our manuscript and limited several claims, which were too extreme in our original manuscript. We have added material to the supplement which provide useful background for linked reads and expand upon some of the finer points of linked reads.

*Glossary.* We believe significant confusion is created by the number of terms in our paper that may be unfamiliar to readers. To reduce confusion we have made our use of novel language more disciplined and added a glossary defining all terms related to Linked-Read sequencing relevant to our paper.

*Read Cloud Sizes.*  We have added a supplementary figure showing the distributions of read cloud sizes in our data.

*Barcoded Reads.* We have added supplementary figures showing basic properties of barcoded reads in our datasets in case these are of interest.

**Point by point response**

**Reviewer 1**

*1) The paper is well-written, and the algorithm is clear, but the explanation around the equations are sometimes confusing. For example, I’d like to know why the authors chose a geometric distribution based on the genome index (i.e., the i in G\_i), or what, precisely, they mean by “conservative parameters” based on the equations they provide? Specifically, I feel that the notation in this section could be improved, and more space could be spent on exposition to motivate the model and make the derived probabilities more clear.*

*Equations*. Reviewer 1 mentioned that they found our use of equations unclear at points. To address this we have simplified our equations where possible, confirmed that the equations do represent what we intended, and have added line by line annotations.

Combined with our simplified use of language we believe that our equations are now much more readable. We have also addressed specific concerns about parameters, geometric distribution, and similar.

*2) The references to NLP and topic modeling early in the manuscript were at times confusing. Specifically, apart from the motivation in developing the graph-based barcode deconvolution algorithm, topic models don’t really seem to come up until section 2.6.1. Here, the authors actually use topic modeling to cluster the enhanced barcodes. Although this issue is relatively minor, I feel the manuscript would be clearer if greater distinctions were made between places where “generative modeling” ideas simply inspired the algorithm (sections 2.2-2.4), and places where topic modeling was actually used. Also, while I realize there may be a difference between communities, I found the use of the term “Graphical Model” (e.g. in the title of section 2.2) particularly confusing, as the authors are discussing a graph-based model, not a probabilistic graphical model.*

*NLP and Topic Modeling.* We have removed language that references our final algorithm as natural language processing and have made it more clear where our algorithm was inspired by natural language processing.

*3) Figure 1 is quite unclear. What does the blue bar at the top of the figure represent? Why are reads from overlapping fragments seemingly tagged with the same barcode (if this is what the circles at the bottom of the figure represent), when the derivation on page 4 argues that this should not happen (or only happen very infrequently). Also, Figure 1 appears not to actually be referenced or explained in the manuscript’s text, making interpretation even more difficult.*

*Figure 1.* We have revised figure 1, added a better legend, and referenced it from the text.

*4) In Figure 2, parts 3 and 4 are a bit confusing, as it seems one is comparing the reads from a given barcode (red) to all other barcodes (purple). The other barcodes should probably each be given their own color, as in the current figure it is not clear that they are distinct.*

*Figure 2.* We have revised figure 2 to remove unnecessary detail and have added improved annotations.

*5) The authors don’t provide any analysis of the performance of their method (e.g., time, memory, and space requirements). Even if there is no other method against which to compare, the computational requirements of the proposed approach should be provided and discussed.*

*Runtime Performance.* We have added a table listing performance requirements for our tool with both datasets and various parameters.

*6) In section 2.2.2 it is not completely clear if the last step of finding connected components happens per barcode read set, or for all reads together? From the text of the paper, it seems that the deconvolution for reads from each barcode is performed independently. Is this the case? That is, is there an “outer loop” over barcodes? This would also seemingly make the corresponding algorithm embarrassingly parallel. This confusion would be further cleared up if the authors adopted a single “pipeline”/”algorithm” overview or figure, as suggested in the next comment.*

*7) I’d suggest the authors gather in a single place (e.g., as a bulleted list, or pseudocode), the set of all of the steps of the proposed pipeline (including the heuristics used for filtering). For example, there are multiple filtering steps for k-mers that are abundant in one barcode and later are shared among many barcodes. Having a “full pipeline” overview would help in this context.*

*Algorithm and Overview.* We have added a cohesive step by step summary of our algorithm in the methods section. We have also added a small ‘refresher’ description at the beginning of the results section.

*8) I couldn’t find supplementary figure 1 to see how the parameters affect Minerva’s performance, which I’m very curious about --- especially the window size and the value of k for minimum sparse hashing. More generally, in sections 2.5 and 4.1, the authors refer to the supplementary material, but it appears that none was uploaded with this submission.*

*Parameters and Performance.* We have added a supplementary figure showing how parameters affect algorithm performance.

*9) I am curious if the authors believe there is a (mathematically tractable) way to calculate the impact of noise on the assumptions behind the algorithm? That is, how string similarity can affect the probability that two overlapped reads come from fragments with different sequence rather than the same (or almost identical). Repeats and near-repeats are the plague of many genomics algorithms, and it would be quite useful to know, mathematically, how they affect the proposed pipeline.*

*Repetitive Sequence and Noise.* We agree that the proposed question is interesting and relevant but we cannot provide a reasonable mathematical model of the effect of homologous sequence on our algorithm and must defer to empirical results. However we have added estimates of the minimum and maximum false positive rates from which communities could be recovered

*10) In the “Enhancement” section, the authors suggest a pair of optional enhancements to the proposed algorithm. However, the don’t use these enhancements to generate the results obtained in the manuscript, and don’t evaluate the practical effect of these enhancements on processing data. Also, no recommendations are given as to when one may (or may not) wish to apply these enhancements. I would recommend the authors either remove this section (as it currently seems somewhat extraneous), or provide more results and discussion on the effect of these proposed enhancements and when to use them.*

*Enhancements.* We have moved this section to the supplement.

**Reviewer 2**

*\* Figure 1: The legend isn't very descriptive and it shows every fragment overlapping (assuming the thin colored lines are fragments). In section 2.1 your model relies on "G as a discrete collection of exactly Ng fragments Fi that do not overlap". What are the little gray boxes under the barcode pools? Some labels would do wonders here.*

*Figure 1.* We have revised figure 1, added a better legend, and referenced it from the text.

*\* Figure 2: If you are going to refer to "Parts" in section 2.2.2, label them on the figure. Again, labels or a key would help. Make sure the text agrees with the legend. For instance the figure says "2) Fragments are sequenced and tagged" but the text says, "dataset is parsed into a set of minimizing kmers (see section 2.4, figure 2 part 2)."*

*Figure 2.* We have revised figure 2 to remove unnecessary detail and have added improved annotations that match the text of the paper.

*\* Table 1: "10\*10^6" Do you mean 1x10^6 or 1x10^7?*

*Table 1.* We have ensured that all numbers are in scientific notation rather than engineering notation or shorthand abbreviation.

*\* Figure 5 is never referenced in the text. I had to look up the accessions to figure out that it's probably based on Dataset 1. There are nine accessions because plasmids are included for the 5 species. That's interesting but it's not discussed anywhere. Actual species labels would be better than accessions*

*Figure 5.* We added an in text reference to figure 5.

\* Figure 6: What do the pink/blue colors mean? taxonomy?

*Figure 6.* We have rebuilt our results for taxonomic promotion and added a table describing our results. This table replaces figure 6. Our new results compare and contrast the use of standard and enhanced read clouds which we believe is the most relevant comparison we can give here.

*2. The central claim of the paper is: "we demonstrate that deconvolved barcoded reads significantly improve downstream results by improving the specificity of taxonomic assignments, and by improving the ability of topic models to identify clusters of related sequences." These two points are somewhat demonstrated in figures 5 and 6, but it's not much, really just isolated examples. For figure 5 it's not a surprise that unprocessed barcode pools don't show any meaningful clustering. You would be better to compare this to some other barcode binning tool or even just binning by tetranucleotide frequency. I get that your method is a new class of methods, but you need to demonstrate why someone should download this. You could do the same for figure 6, does tetranucleotide binning rescue the taxonomic classification of E. coli reads? You can't claim "significant improvement" without a fair comparison.*

*Downstream results.* We did not intend for this to be interpreted as the major impact of our paper. While we believe that barcode deconvolution has the potential to improve downstream results Reviewer 2 is quite correct to state that we have not successfully made that case. Correspondingly we have weakened our language surrounding downstream results and made our downstream results section more thorough.

*Minor comments:*

*1. Section 1.2, first paragraph needs some references.*

*2. Also in section 1.2 "Compared to long-read sequencing, linked-reads can be used to sequence samples far more deeply for the same amount of money..." This sentence is a little confusing because you state above that linked reads only fractionally cover DNA fragments. I \*think\* you're saying that \*if\* you spent the same amount of money as conventional short-read sequencing, it would be deeper but you appear to be contradicting yourselves.*

*3. Section 1.3 has a symbol heavy equation "D(i)-D(j)..." that you summarize by saying, "The barcode deconvolution problem for a set of barcodes is the problem for a single barcode applied to every barcode in the set." The text below implies, perhaps unintentionally, that deconvolving a barcode can be done in isolation when, in fact, you are using a graph of all of the data to solve the problem.*

*4. Section 2.5 "Failed Models" will likely be on interest to some but it belongs in the supplement.*

*5. Section 2.6.1 You don't define "enhanced barcodes" until section 4.1*

*6. Section 3 The mock communities are not described very well. How were they made? Were they normalized to the same number of cells per organism or same concentration of DNA per organism? Maybe they were staggered concentrations? Where did the organisms come from? ATCC numbers? Why is there human contamination?*

*7. Pg 9 "evenly distributed over 3M barcodes", I'm not sure what you mean by 3M.*

*8. Bottom of page 9, "We used LDA to project deconvolved and standard barcodes". Do you mean "enhanced vs standard". Be consistent. There is a lot of specialized language in this manuscript so you need to be very clear throughout.*

*9. Section 3 paragraph 1, you incorrectly spell or abbreviate at least 4 microbe's names. Also section 4.3 has E. Coli without italics.*

*10. Italics for latin "a priori" and "de novo" throughout. No hyphens.*

*11. Section 5, "The current version of Minerva provides reasonable performance but can still up to 100 cpu-hours to deconvolve some large-scale datasets." What is large? If you are going to talk about performance, be specific.*

*12. "de Bruijn" isn't hyphenated, but it does have a capital B.*

*13. References to "fresh" sequencing data in Availability and Acknowledgments. Find a better word. Also, why are the data sets upon request only? Shouldn't they be deposited at NCBI? They would be valuable for others trying to replicate or improve on this work.*

*Minor comments.*

1. We have clarified that the use of linked-reads in metagenomics is largely speculative.
2. This has been clarified.
3. This text has been clarified and the equation has been presented with more annotation.
4. This has been moved to the supplement
5. We have changed this term to enhanced read clouds and added a glossary
6. We have described why we chose mock communities and their role as standards
7. We have changed all numbers to scientific notation
8. We have made our language consistent
9. We have fixed the spellings of microbial names.
10. We have added italics to all latin phrases.
11. We have removed vague references to large datasets and timing
12. ‘de Bruijn’ is capitalized correctly
13. We have made the data available for download.

**Reviewer 3.**

*- As written, this manuscript is not accessible to the Genome Research audience. It includes some complex ideas that would benefit from simple explanations and intuitive examples. The narrative is confusing in many places (specific examples below).*

*Inaccessibility.* We have clarified language and equations in the hope of making our paper as accessible as possible.

*- Proof read the manuscript for typos and repetitive text.*

*Proof Reading.* We have now proof read the new manuscript several times.

*- Mathematical formalism is useful for communicating the ideas in this manuscript, but please check that all terms in the equations are carefully defined and notation is consistent.*

*Equations*. We have simplified our equations where possible, confirmed that the equations do represent what we intended, and have added line by line annotations.

*- Inclusion of approaches that did not work is appreciated.*

*Failed Approaches.* We appreciate reviewer 3’s comment and note that we have moved this section to the supplement in line with reviewer 2’s comment.

*- #1 issue: The deconvolution problem is not clear. In section 1.3, I am perplexed about what the unobserved class r\_i for read i represents and how it is different from the fragment f\_i for read i. The r\_i and the f\_i are both defined as representing fragments. Might r\_i actually represent the source genome of read i? If so, a lot of what follows makes more sense to me. But I might be confused…*

*- The specific goal of Minerva should be stated in words up front and then linked to the math. If I am right about R representing genomes, it appears that the method clusters barcodes based on shared k-mers to then group together reads from the same genome across barcodes. Is this right? If so, you could say that you are trying to infer the genome for each read (i.e., infer R from F). And in the methods section, you can state that the clusters output by Minerva are an estimate of the genome sources (i.e., R).*

*Deconvolution Problem.* We have added a plain text description of the problem, clarified the meanings of each symbol in our equations, and added line by line annotations to our equations.

*- Once it is clear what the deconvolution problem is, the motivation for why barcode deconvolution is hard in metagenomics (intro, section 1.3, and throughout) needs work. Several of the challenges from sequencing of a single genome (small or structurally rearranged genomes that violate the assumption that fragments do not overlap and hence make mapping-based deconvolution hard) do not seem to apply to metagenomics, where the total genome length in a complex community is not small and most organisms do not have a genome any way. Lack of reference genomes is definitely an issue, and perhaps this is the best motivation. If there is also an issue of fragments from the same genome getting the same barcode, some data supporting this claim would be helpful. This seems unlikely in diverse communities unless they have a very skewed species abundance distribution.*

*Deconvolution in Metagenomics.*  We have clarified that the primary issue of barcode deconvolution in metagenomics is the lack of known reference genomes.

*- Intro: Provide a citation for there being 2-20 fragments per barcode.*

*Fragments per Barcode.* We have clarified that this range is observed in our data.

*- Intro and motivation sections: It would be nice to say more about how the genome clusters output by Minerva can be used differently from other metagenome analysis outputs. The applications help, but it would be great to spell this out.*

*Genome Clusters.*  We believe that this criticism arose from a poorly worded section in our manuscript that we have since changed. It does not necessarily reflect the goal of our algorithm.

*- P.3: The statement that “reads from the same fragment will overlap with similar barcodes” seems central to the motivation and statement of the problem Minerva solves. But it is not clear what “overlap with similar barcodes” means.*

*Barcode Overlap.* We have significantly clarified and increased the language surrounding this section.

*- P.3: Provide some justification for partitioning genomes rather than drawing random fragments, which is what happens in sequencing.*

*Genome Partition.* We have clarified that this is a conceit of our model to make the mathematics simpler.

*- P.4: It is not clear why these are conservative or -importantly- realistic parameters.*

*Conservative Parameters.* We have clarified why these parameters are conservative an why they are realistic.

*- P.4: Explain why it is important that two barcodes sharing exactly one fragment is more likely than two barcodes sharing two or more fragments.*

*Number of Fragments.* We have clarified why it is important that read clouds share one fragment much more often than more than one fragment.

*- Fig 2: Not clear which parts the numbers 1-6 refer to. Explain colors and shapes.*

*Figure 2.* We have revised figure 2 to remove unnecessary detail and have added improved annotations that match the text of the paper.

*- P.5: What effect does dropping kmers with frequency equal to 10 times the mean have on performance quantitatively, and how did you arrive at 10 times the mean?*

*10 Times Mean.*  We have clarified that 10 times the mean is an arbitrary threshold but that is based off of our mathematical model in principle.

*- P.6: Where does the typical number of 100 nodes come from? And which nodes are being referenced (reads, barcodes, something else)?*

*100 Nodes.* We have clarified that this number is observed in practice and refers to all nodes in the graph.

*- It is not clear if the bound in section 2.3 depends on the simple model in section 1.4.*

*Boundary.*  We have clarified that this boundary is not specifically related to our mathematical model.

*- Section 2.4: Elaborate on how hashing affects accuracy and compute resources for Minerva.*

*Hashing.* We have clarified why and how hashing improves the performance of Minerva.

*- Section 3: Mock communities are a good way to benchmark. The ones analyzed are pretty simple in terms of richness and evenness compared to real data. What is lost and gained by using them? Can more realistic performance be assessed (e.g., computationally) or predicted?*

*Mock Communities.* Currently we are not aware of any tools to simulate linked-read data for metagenomics. Rather than rely on an unrealistic computational simulation we used real sequencing data from commercially available mock communities. The Zymo communties we used are standard positive controls (recommended by IMMSA) and should facilitate comparison with other parts of the field. We agree that the mock communities used are simple however we are unable to produce a more complex community. We have clarified these points in our manuscript.

*- Section 4.1: Do you mean groups represent genomes? Not clear how the same exact fragment ends up in multiple barcodes even in a simple community. Also, give more detail of what not-enhanced and enhanced barcodes are.*

*Section 4.1.* This was essentially a typo and we have corrected it. We have also added a glossary that clarifies many of the terms we use.

*- Section 4.2: I don’t fully understand what you are doing in this section.*

*Section 4.2* We have clarified how we are using LDA to cluster read clouds.