Dear Editors,

We would like to extend a sincere thankyou to both the editor and the reviewers for their thorough and constructive comments.

In light of the received feedback, we took the time to extend the simulation to include a deeper level of HiC sampling, which contributed to and made clearer the prevailing trend in clustering performance for the selected algorithms. Recognising the time and effort in reproducing our work from scratch, we have also organised and published the complete supporting simulated dataset.

In the proceeding rebuttal, all of our responses are in blue.

Regards,

Matthew DeMaere and Aaron Darling

Editor's Comments

I have received two thorough reviews of your paper and both reviewers agree that your paper is of interest to our readership and offers a solid study on an interesting topic. However, both reviewers identify a number of issues that you will need to deal with before your paper can move forward. Please carefully consider these reviews. I concur with their general assessment and details of suggestions. Good luck in your revision.

Reviewer 1 (Anonymous)

Basic reporting

In general, the paper is well written. It is self-contained and gives a comprehensive review of the relevant methods and previous work.

The most important Figures 4 and 6 are placed at the very end of the paper, rather than close to relevant paragraph making them slightly harder to follow. Also, Figure 6 is overloaded with data, making it hard to interpret and match to the textual description.

We have let the placement of figures be dictated by the latex engine, which often places near page-sized figures at the end of manuscripts. We're assuming that in final production these would be set by the publisher. Figure 6 does require careful study of the reader for interpretation, but we think it better to present this more complete set of results than to further reduce the results described in the paper.

Experimental design

The paper studies the performance of different graph clustering algorithms applied to the problem of binning metagenomic assembly contigs based on 3C data linkage.

I think that the described research falls into the scope of the journal.

As discussed in the paper, most previous works on metagenomic binning with 3C (HiC) data have been using graph clustering approaches. At the same time, performance of those approaches has not been thoroughly assessed in presence of closely-related genomes. This study fills this knowledge gap by assessing performance of different clustering algorithms on extensive set of simulated datasets. Thus I find discussed problem is relevant and meaningful.

Validity of the findings

1. It is not clear why the authors sometimes implement 3rd party algorithms themselves, while there are readily available implementations. For example, on line 208, they mention their implementation of OClustR algorithm (P´erez-Su´arez et al., 2013) in Python, while Perez and Suarez provide implementation in C++. Why not just use this implementation?

We are strong advocates for the re-use and extension of existing open-source software whenever possible and would have always taken available codebases over reimplementation. To that end, MCL and SR-MCL were obtained from their repositories. In the case of Louvain, our implementation is to only provide a CLI and I/O operations for an already published 3rd party module (community). In the case of OClustR, we did contact the authors but were denied access to both source and binary, citing commercial secrecy. Instead, the authors offered only to assist us by answering questions when we undertook our own implementation. In the time since that correspondence with Perez-Suarez et al, we noted a subsequent publication from the same (or related) group regarding a GPU accelerated implementation. Unfortunately, no code or binary was made available with that publication either. We did not pursue it by email a second time. In other areas, implementation was done only when required and with attention to employing preexisting mature modules where possible.

There were some unavoidable sunk costs. In the time since we initiated our work, it is possible that some implementations have appeared. Though we have every respect for the open-source ecosystem, code appearing on hosting services such as Github, may not have been published with the explicit intention for use in “production” (for lack of a better term).

As testing and validation is ad-hoc and voluntary in such publication processes, the quality standard is often uncertain and frequently inferred simply from public uptake. Therefore proper diligence requires time to be set aside to validate unfamiliar, not-widely used code. In the end, we do not consider the development of our own implementations of these methods an achievement, but rather a necessity.

2. As far as I understood, the raw data has not been made available, but the instructions to reproduce the simulations are given in the paper. Unfortunately, relevant scripts have not been provided to the reviewers.

The resulting dataset has now been made publicly accessible, in addition to the github repository. (http://doi.org/10.4225/59/57b0f832e013c)

On reflection, we decided the best course of action for assisting other researchers in reproducing our work was to provide the data. That said, work is currently underway on a new, streamlined simulation and testing system that will make generating a similar data set much easier and will enable researcher to simply quote the governing parameter set and random seed, but is outside our intended scope here.

3. For a fixed set of parameters, only a single dataset has been generated. Having more replicates should increase the robustness of the findings.

Although the work did not generate explicit replicates, we would argue that the two phylogenies operate well enough as proxies for one another, so long as no unduly precise conclusions are drawn from resulting observations. We avoided drawing strong conclusions on individual simulation points, and agree that to do so would require replicates to estimate variance and confidence intervals for simulation results.

Our aim was to demonstrate gross behaviour as a function of wide variation in experimental parameters and we did not attempt to take fine measurements from these observations, which would be more susceptible to stochastic aberration. Across the entire sweep, the discussed large-scale trends are, more so than not, smoothly varying. Additionally, the correlation tests among observed/measured parameters were calculated over many elements (e.g. the entire set of simulations).

Although not shown in the manuscript, we did perform a subset of the total sweep with varying-seed replicates and it reproduced the trends seen in the manuscript.

In general, the findings confirm reasonable expectations, but show interesting particular boundaries on the applicability of different popular methods.

Comments for the Author

Major comments:

1. The question of “sufficient” 3C data depth was left beyond the scope. Would 10^6 HiC paired-reads give even better results for soft-clustering approach? When should the saturation be expected?

We have extended the sweep to include this depth (n3c=1M read-pairs). The additional step shows that saturation has been achieved with respect to the hard-clustering algorithms (MCL, Louvain-hard). For the remaining three soft-clustering algorithms, a beneficial

response is most clearly evident in Louvain-soft, with that of OClustR being smaller. For LS, this is largely because of improvement in Recall, despite Precision suffering slightly.

2. Related important question: if the perfect linkage data was given, how would clustering algorithms perform? It seems that testing algorithms in a perfect conditions first is crucial for understanding of their limitations!

We agree this is an interesting question, however in the context of genomics it is one that is impossible to answer. The process of contig assembly does not preserve information about the genomic source of reads that become coassembled into a contig. Further, the process that leads to coassembly is itself complex, with unpredictable fuzzy outcomes that vary with genome assembler. Therefore it is particularly difficult to know, beyond a practical precision limit, which contigs belong to which strains, and therefore impossible to create a graph with perfect linkage.

The question is a topic of research in pure graph-clustering research, enabled by tools such as the LFR benchmarks (Lancichinetti 2008).

* Lancichinetti A, Fortunato S, Radicchi F 2008. Benchmark graphs for testing community detection algorithms. *Physical review E*.

3. It seems that “ladder” phylogeny is largely useless! Interpretation of corresponding results is much harder and the community is still far too simple to be presented as an approximation to the real case.

We chose to include a second phylogenetic topology out of concern for the objection that in not doing so, we have made the simulation to simplistic on all accounts. It also acted to provide evidence of robust trends as a function of evolutionary divergence. A futher utility of this phylogeny is that the transition across the species/strain boundary is more progressive with respect to the entire community, the effect of which can be seen the the less abrupt transitions in R_{B3} and subsequently F_{B3}. Including two phylogenies and two abundance profiles goes to the question of replicates above, with very similar large-scale trends are apparent in each.

4. In general, the impressive complexity of the methods and measures used (definitely a strong side of this work) seems to be in disproportion with the simplicity of the setting which is being analysed. 4-strain mixtures used in the study appear to be a significant oversimplification of strain-mixtures in real metagenomic communities. Would the increase in number of strains change the result?

Good question. These phylogenies were chosen precisely for that simplicity and the ease of interpreting the behaviour of algorithms without the inevitable confounding effects of specifying a community complexity verging on realism. This consideration involves not just assessing clustering algorithms, but the methods used in steps internal to the pipeline. Now armed with a more mature understanding, more realistic and complex simulations could be done but would be a substantial body of work, better presented as a separate publication.

5. Principal components analysis interpretation is a bit cryptic and looks like an overkill.

Effort has been made to improve readability of the figure.

The PCA biplot was included so as to jointly summarize all parameters and the resulting measures from all simulations of the star phylogeny. It provides a visual exploratory guide for correlations among parameters and not just those on which we selectively mention by statistical testing. We feel this is important to include rather than rely on a small number of slices across the sweep such as Figure 6. In particular, it is possible to see the effect of increasing HiC read-depth on the performance of the five algorithms -- and that Louvain-soft responds more noticeably than the other four.

6. Moreover, increase in HiC depth also leads to increased graph complexity!

In fact, as can be seen in the biplot of figure 5, increase in HiC depth (n3c) is nearly orthogonal to our measure of graph complexity (H_L). Although the number of edges (*size*) does increase with increasing HiC depth, this does not greatly affect graph complexity H_L.

7. Should not weight of the edges in the contig graph be normalized by the contig lengths?

Normalization of edges is indeed often done with experimental data, however as was outlined in the discussion under limitations, we found that for our simulation data normalization was not necessary and even a hindrance depending on the clustering algorithm, community and sampling effort. Further, after trialing a number of published normalization strategies on what experimental data has been published to date, it is not clear that these simple relations between length or number of RE sites consistently leads to better outcomes.

8. The paper refer to clustering algorithms with a given number of clusters as supervised algorithms. Traditionally, supervised learning means that there is some training data set, is that the case for their studies? Although for K-means clustering determining the value of K can be an issue, usually it refers to unsupervised or semi-supervised learning algorithms.

Revised. Wording has been updated.

Minor comments and misprints:

1. Introduction states that “Computational approaches to genome assembly would have little success if not for the … assumption that a very high degree of sequence identity (>95% ANI) implies a common chromosomal origin.” The sentence is not clear. Reviewer is unaware of genome assemblers that use this assumption.

Revised

2. On line 98, it should be “been” rather than “be”.

Fixed

3. Beginning of third page states “including varying degrees of evolutionary divergence around the species boundary”, while “species boundary” is not explained up until much later in the text. Reader would benefit from explaining it here.

Revised text and added a reference to the figure.

4. Logical formula on line 128 lacks some quantifiers and uses \subset in place of \in. Logical formula on line 130 is a tautology (uses \supset in place of \ne).

Rephrased the set notation at 128 in terms of clusters only as referencing the objects leads to an unwieldy definition. Rephrased overlap on 130 in terms of cardinality.

5. The paper states: "For instance, in the metagenomic analysis of closely related species or strains, the tendency of the highly conserved core genome to co-assemble into single contigs while the more distinct accessory genomes tend not to, implies that a 1-to-1 correspondence of cluster-to-genome is not possible and an overlapping model is required."

   Here, it is not clear how contigs are related to clusters (and what are objects being clustered at first hand), and how an overlapping model would help.

   Also, "more distinct accessory REGIONS" rather than "genomes" would be more appropriate, otherwise meaning of “conserved core genome” needs to be specified.

   Also, "core genome(s)" and "single contig(s)" should be in the same (singular or plural) form.

Revised, thanks for the helpful suggestions.

6. On line 232, notation “|x|” for the total genome length is weird, since x denotes a specific position.

Revised notation.

7. On line 240, “trans” and “cis” read-pairs are introduced relatively to set of contigs, while these notations have already been used before in a different context.

Unintended misuse of terms removed.

8. Have the authors tried to work with real E.coli strains? Many finished reference genomes are available at various evolutionary distances.

We have not attempted to approximate our simulated communities using real *E.coli* strains from the pool of available genomes. Although this might further support our observations, we feel time would be better spent considering more complex communities and real metagenomic samples. However, including this as an alternative run method in a future pipeline is intriguing.

9. The “1/e” community abundance profile is not explained.

Added an explanation.

10. On line 402, should it be “r=0.61” ?

You are correct, revised.

11. On lines 404-407, increase in edge complexity is stated as a bad feature, which is misleading, since edges give the information for the clustering! At the same time, while it is stated that “increased 3C sampling depth can significantly improve the

quality of clustering solutions", it is not explained why it should help if the edge complexity stopped increasing?

Revised this section. It was poorly worded after some strong efforts to revise before submission.

12. On lines 458-459, alternative "more technically complicated HiC protocol" is mentioned but neither referenced nor explained.

Added citation to method. HiC's application to metagenomics was mentioned in the introduction with citations.

13. What is "graphical order" in "Graphical complexity (HL) exhibits a similar turning point to L50 and for the 3C contig graph is dominated by graphical order"?

Made this clearer.

14. Addition of "sanity-check" dataset with distant genomes and inter-chromosome read-pair background is highly suggested.

As the simulated genomes are mono-chromosomal, inter-chromosomal associations would mean inter-cellular associations. In published works on real-world HiC experiments (Beitel 2014), it has been shown that inter-cellular associations occur at a rate 10^2 – 10^3 times less frequently than intra-chromosomal and were attributed to random ligation events post-lysis. We address this issue in discussion, where we feel that the process amounts to an weak background in our simulated communities and does not present a significant perturbation.

A thorough examination of the effects of noise on downstream analysis of HiC/3C sequencing data does remain an interesting and substantial topic in and of itself. We would suggest this would be better placed in manuscript of its own.

i. Beitel CW, Lang JM, Korf IF, Michelmore RW, Eisen JA, Darling AE 2014. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* 2:e415. DOI: 10.7717/peerj.415.

Overall, this is an interesting work, clearly showing the limitations of simple graph clustering techniques to the problem of deconvolving microbial communities. But the final message:

"We recommend that future work focus on the application of recent advances soft-clustering methods"

sounds questionable to me. After reading the paper, I would argue that better deconvolution techniques should be developed, rather than staying limited to the general-purpose graph clustering algorithms! It seems that graph clustering techniques will always give many precision errors. For example, imagine conservative regions and 2 strains with long single strain-specific region in each. They will be very well connected through conservative regions, but two long fragments not connected by linkage info should not be placed into the same cluster at all costs!

In the context of 3C-contig graphs and using a conservative assembler, the most likely outcome from the provided example would be the creation of breakpoints at the confounding

junctions. The junctions would then lead to multiple contigs where conserved (C) and strain-specific (SS) regions are sequestered separately.

True class membership would then be multivalued for C contigs and singular for SS contigs. This scenario, where objects may be a member of one or more classes, is precisely the problem soft-clustering algorithms intend to solve and so we believe our choice is a good one. Though not all soft-clustering algorithms are graph-based, for HiC data the representation is a natural one.

Future advances will likely exploit finer graph granularity than that provided by whole contigs. We chose to focus on simple 3C-contig graphs for interpretability and computational feasibility.

Reviewer 2 (Ben Woodcroft)

Basic reporting

In general the introduction is adequate and the methods section is for the most part clear, providing sufficient detail. However, I found the results section to be insufficiently contextualised. While as a reviewer I had time to read the manuscript from start to finish, I do not expect that the average reader has such a luxury. Thus, the results section should briefly explain the experimental design before presenting the results of those experiments and their interpretation.

A description of the experimental design has been added to Results.

In general the figures and figure legends could be more clearly presented. For instance in Figure 5 the labels are often unreadable, and it is not immediately clear what “individuals” and “variables” refer to. In particular:

Revised the caption for Fig5.

1. Figure 3: The acronym “BL” should be spelled out.

Fixed, thanks for catching this.

2. Figure 4 y-axis labels should have units (where applicable).

Revised figure for improved clarity. Most axes are unitless, but I’ve added units where applicable.

3. Figure 4c: The dashed lines should be included in the legend. The dashed lines are suggested in the figure to be first order derivatives of the solid lines if I understand correctly. However, this appears to be incorrect as the slope of the solid lines is clearly positive around 97% ANIb where the “first derivative” is graphed as negative.

The figure has been revised. Notably, A sign inversion in dS was missed when switching from using a_{BL} to ANI_b for x. Thank you for catching this error.

4. Figure 5 & 6. It is unclear what “nhic” refers to. I would guess this is the same as “n3c” but this should be clarified, and acronyms used more sparsely in the figure legends.

This error was previously noticed immediately after submission and was corrected in a preprint revision.

The repository of code at https://github.com/koadman/proxigenomics should be mentioned in the main text of the manuscript. I was only able to find reference to it in the Declarations section. The focus of the manuscript is the interpretation of the results of this code as opposed to the code itself, but nonetheless it would be helpful if readers were instructed as to how central parts of the study relate to the code base. Also, it would be helpful if the base

directory of the repository contained a README which outlined the broad structure of the repository as a whole.

As PeerJ included a facility for declaring the code repository, we removed the reference from the body of the text. I have now reintroduced an explicit note.

Further minor comments:

1. * 46: cell lysis is not the only process that happens during the DNA purification step.

Changed line to read (during the process of DNA purification) since we were not intending to imply that this was the only detail of purification, just the one relevant to discussion.

2. * 52: Assembly algorithms do not require this assumption to be uniformly true as this sentence implies.

The last remark has been removed.

3. * 56: Assembly is not generally conceptualised in terms of a “system” in this sentence, making this sentence unclear.

The next statement attempts to clarify what an underdetermined system is, but I have now replaced “system” with “inverse problem”. The use of this term in genomics goes back to early publications on the human genome (Venter et al 2001 10.1126/science.1058040) and continues today (Myers 2016 0.1515/itit-2015-0047).

4. * Paragraph starting 69: This should be better structured, introducing the basic idea of the family of methods before listing the different sub-types.

Restructured this section as suggested.

5. * 73: “Thus” does not make sense here.

Revised

6. * 84: The implication here is that one read comes from each side of proximity ligation. As in the discussion (line 493), this is not true for all cases. While this simplification could be considered appropriate for the manuscript as a whole, it should be explained in the introduction.

The following text has been added.

> As with any real experimental process, the generation HiC/3C read-sets is imperfect. Three complications to downstream signal processing are: self-self re-ligations which effectively produce local read-pairs, chimeric read-throughs which span the ligation junction and contain sequence from both ends, and spurious read-pairs involving non-proximity ligation products. Though not insurmountable when integrating HiC/3C data with that of conventional sequencing, these flawed products do at the very least represent a loss of efficiency in generating informative proximity ligation read-pairs.

7. * 90: It is unclear whether the author suggest deconvolution of raw metagenome reads or contigs/scaffolds. Certainly, these techniques cannot deconvolute the actual cells themselves (as I believe the authors do not intend to imply).

Revised sentence to be more explicit about what we propose to deconvoluted with HiC/3C.

8. * 116 and throughout: the term “contig graph” is not inappropriate, but for the sake of avoiding confusion with the more standard meaning of this term in genome assembly I suggest an alternative name be used.

Revised all uses to read “3C-contig graph”.

9. * 129: “objects allowed”: missing “are”

Fixed

10. * 140: “would be” should be replaced by “are”

Fixed

11. * 143: It seems self-evident that validation measures not be primarily concerned with algorithmic (time) complexity of the algorithms but instead their utility in providing a solution to the problem at hand. Thus the introductory sentences are not needed.

We agree that in principle we should be unconcerned with the compute time for validation, however in practice it is a critical detail. Validation measures within the Bcubed family can have poor time complexity, i.e. CICE-BCubed* at $O(n^3 \log n)$, where n is the number of nodes. In our work, it was found that validation took longer than the act of clustering for larger graphs. In particular validation measures which first must match class-to-cluster prior to validation, compute time required for validation made their application impractical.

*Rosales-Méndez H, Ramírez-Cruz Y 2013. CICE-BCubed: A new evaluation measure for overlapping clustering algorithms. *Iberoamerican Congress on Pattern ….*

12. * 214: This paragraph is confused and overly long if the main point of the paragraph is simply that 3C reads need to be mapped rather than extracted from the de-Bruijn graph.

Corrected. A bit of entropy creeped in during early revision.

13. * 281: It should be made more explicit that the 3C reads were simulated using the full genomic sequences rather than the contigs. I also suggest that the read simulation is sufficiently important that it should be included in its own section toward the beginning of the Methods rather than in the catch-all “Pipeline Design” subsection.

Revised. A brief description of the HiC/3C simulator was made a separate subsection of Methods.

14. * 362: The HL graph complexity measure is not given a name (beyond “HL” and “non-parametric entropy”) and I could not find HL (by that name) discussed in either of the two references given. It may be helpful if the original article proposing the measure be cited as opposed to articles which discuss them.

My mistake, the symbology was taken from a reference that I failed to cite, which I have now added. That said, the equation is not explicitly written in the body of this reference as it deals with a family of entropic measures differing by the choice of what they refer to as descriptors (e.g. Laplacian or Adjacency matrices).

H_L can be found in Table 1 p157 of:

*Dehmer M, Sivakumar L 2012. Uniquely Discriminating Molecular Structures Using Novel Eigenvalue—Based Descriptors. *Match-Communications in ….*

15. * 374 and elsewhere in the Results: when discussing the results of experiments, it should be done in past tense rather than present tense.

Revised

16. * 414: p-values should be reported as e.g. “<1x10^92” rather than “2.77x10^-93” as the exact values are not especially relevant.

The precision of reported p-values has been reduced as suggested.

17. * 426: The first paragraphs of the discussion appear to be more like results than discussion.

As the reviewer previously suggested for another section, the first paragraph of the results was intended to aid the time-poor reader.

Being qualitative, these are less-so hard results and more-so an aid to discussion. We do not feel that placing it further away, within results, would not make the manuscript clearer. I have tried to reduce that which detracts from the point of discussion.

18. * 545: “increase” should be plural.

Revised.

Experimental design

The approaches presented in the article appear sound. My only concern is that only the uniformly distributed community structure was used to assess the clustering algorithms. Uniformly distributions of microbial community abundances are very rare in nature, making interpretation of these data difficult if these clustering methods are to be used in the context of 3C data derived from real microbial communities.

Validity of the findings

The largest concern I have with the approach is that no error was incorporated into the simulation of 3C pairs. Not including erroneous links between contigs risks making the simulation overly simplistic. While I am unaware of any simulator for metagenomic 3C data, analysis of clustering algorithms without error may hide complexities that hinder algorithm performance on real data. For instance, while not directly affecting any of the clustering

algorithms, the graph complexity measure would be affected by the presence of spurious edges in the adjacency matrix. If as the authors claim in the discussion, “a simple low frequency threshold removal … could in principle solve the problem”, then this approach could be implemented to alleviate these concerns.

We agree that an accurate error model is an important feature to eventually include in the longer term, particularly if interrogating finer-scale objects than the relatively simple 3C-contig graphs.

Like early read simulators of WGS used in assembler development, idealistically simulated results do not necessarily prevent early advances in analysis or interpretation, so long as the researcher is careful to keep this in mind. The aspect we are attempting to highlight is how the transition from species to below strains may impact commonly deployed methods and the assumptions therein. How this affects assembly outcome (WGS reads were modelled with error) and what this intermediate result means for downstream clustering.

We have been careful to qualify our interpretation of the results and discussion in light of the simplification.

The discussion and conclusion sections tend to imply that the 3C clustering/binning problem can be conceptualised as a graph-based problem, but reducing the data to fit this model is not without information loss. For instance, mismatches between the sequences of the 3C read pairs and the contig to which they map might be used to infer that the contig is of mixed origin. This information is destroyed by converting the data into a collection of nodes and weighted edges. As I imagine the authors are aware, there are a number of other information sources that do not neatly fit into a graph-based framework. This does not mean that graph-based methodologies developed in the article are without value, but that they should be better contextualised in the discussion.

The representation (3C-contig graph) was chosen as it has previously been used in studies of the applicability of 3C data to metagenomics and is simple in nature. We strongly agree with the reviewer that other, more than likely finer-scale, representations of the information will prove more effective in advancing metagenomic HiC/3C analysis.

We have now emphasised the simplicity of the chosen model in the *Conclusion* and *Representation* subsection of *Methods and Materials*.