Cell Systems

Transparent Peer Review Record

Benchmarked approaches for reconstructions of in vitro cell lineages and in silico models of Caenorhabditis elegans, Mus musculus developmental trees

Wuming Gong, Alejandro Granados, Jingyuan Hu, Matthew G Jones, Ofir Raz, Irepan Salvador-Martínez, Hanrui Zhang, Ke-Huan K. Chow, Il-Youp Kwak, Renata Retkute, Alidivinas Prusokas, Augustinas Prusokas, Alex Khodaverdian, Richard Zhang, Suhas Rao, Robert Wang, Phil Rennert,Vangala G. Saipradeep, Naveen Sivadasan, Aditya Rao, Thomas Joseph, Rajgopal Srinivasan, Jiajie Peng, Lu Han, Xuequn Shang, Daniel J. Garry, Thomas Yu, Verena Chung, Michael Mason, Zhandong Liu, Yuanfang Guan, Nir Yosef, Jay Shendure, Maximilian J. Telford, Ehud Shapiro, Michael B. Elowitz, Pablo Meyer

Summary

Initial Submission: Received Aug. 21, 2020
Scientific editor: Quincey Justman, Ph.D.

First round of review: Number of reviewers: Three
Three confidential, zero signed
Revision invited Oct. 14, 2020
Major changes anticipated
Revision received Feb. 1, 2021

Second round of review: Number of reviewers: Two
Two original, zero new
Two confidential, zero signed
Revision invited Mar. 9, 2021
Accepted May 11, 2021

Data freely available: Yes
Code freely available: Yes

Cell Systems

Transparent Peer Review Record

Editorial decision letter with reviewers' comments, first round of review

Dear Pablo,

Thanks for your patience during an unusually long review process. The reviews of both papers are back and I've appended them below. Reviewers 2 and 3 reviewed both papers, although their order is reversed (sorry about that.)

Two things are clear to me from the reviewers' comments. The first is that the two papers should be knitted together because neither is quite standing on its own. The second is that these papers aren't quite research articles (one doesn't have or need a Methods section); this is especially true given that, if I'm remembering properly, nearly all of the individual entries will become stand-alone preprints or papers, if they haven't already. My solution to both problems is to join the two papers into a single larger piece that retains the current tone and content. I can help you with the joining. The final piece will be will be called a "Synthesis" and can borrow writing techniques from both Review/Perspective articles and normal research articles. I have ideas about how to do this smoothly and effectively.

Let's talk briefly over the phone/Zoom to make sure we're on the same page. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later. I've sent an identical decision for the other manuscript; you should have received it a few minutes ago.

I look forward to seeing your revised, combined manuscript.

All the best,
Quincey

Quincey Justman, Ph.D.
Editor-in-Chief, Cell Systems

Reviewers' comments, manuscript 1:

Reviewer #1: This paper describes a recent DREAM challenge regarding cell lineage reconstruction, benchmarking nine participating methods. The interesting aspect of this paper is the use of a recording technology (intMEMOIR) that enables one to reconstruct cell lineages and mutation status of cells in real time. This paper might be of interest to the readership of this journal if the following concerns are addressed.

1. Relationship of experimental technology used in this challenge and lineage reconstruction tasks that occur in practice.

The authors constructed traced lineages and mutations of 106 experiments using the intMEMOIR system. Specifically, they started each experiment with a single cell with 10 loci that may mutate into one of two states: (1) deletion or (2) inversion. Mutations are irreversible (although this is only stated in Figure 3 caption). It is not clear to me how this experimental setup relates to cell lineage reconstruction tasks that occur in practice. The first paragraph of the introduction mentions lineage reconstruction for understanding development. This needs to be expanded on in several ways. First, other important applications include cancer phylogeny reconstruction, somatic mosaiscm in the brain, clonal heterogeneity in T cells, etc. These must be described. Second, the input data must be described as well as limitations. For instance, data can differ in omic type (RNA, DNA, methylation, etc.), sequencing technology (single-cell vs bulk, long reads vs. short reads, or FISH), types of mutations (single-nucleotide variant, copy-number aberrations, structural variants), etc. I encourage the authors to describe how their experimental design fits with lineage reconstruction tasks that occur in practice. Specifically, I'd like to know:

* How applicable are developed methods to these general settings?
* Are they only applicable to this data?
* What are some key ideas/best practices that carry over.

2. Describe training and test data more carefully

My other main concern is that the description of the training and test data must be improved considerably. I spent a fair amount of time to understand key details that are not described in the main text.

* Figure 3 caption states that edits are irreversible. This needs to be stated in the main text. This is a key restriction.

* Please add all 106 trees to supplement.

* Please provide a table with key statistics of the input, describing the number of leaves, and the number of unique leaves (in terms of genotype) for each experiment.

* Do you measure mutation status of intermediary cells (e.g. internal nodes in the tree)?

* Do you measure time of occurrence of each node (or leaf) in the tree?

* What about the last 24 hours? Do new mutations occur? Is lineage only traced in the first 36 hours?

* How was the split in training and test performed? Randomly? Stratified in terms of tree size? Please explain.

* Did participants have access to the test data?

* Why is there variability in number of leaves?

* How does intMEMOIR build upon MEMOIR?

* How large are the altered regions (in terms of base pairs)?

* Figure 2: define large vs. small (how many leaves).

3. Ambiguity in cell labels?

My final main concern is about ambiguity in genotypes of leaves. Fig S3 suggests that there may be multiple leaves that have identical genotypes. I'd like to know whether their are genotypes that are common to more than two leaves. If this is the case then it is not fair to use RF or triplet distance on all leaves, as there will be exist multiple trees that are identical when removing this type of ambiguity. Please describe.

Minor:

* Line 130: readout => read out
* Line 138: was splitted => was split

Reviewer #2: Granados et al. supplements the results described in the other manuscript to report outcome from the Allen Institute lineage reconstruction DREAM challenge using the intMEMOIR recording data, a synthetic imaging-readable lineage tracing technology. I have the following major concerns of this work. I generally have the same serious concerns of this manuscript as the other one reporting this DREAM challenge outcome. I list some of them below.

- I strongly think this manuscript should be combined with the other one to present an overall report of this DREAM challenge.

- Similar to the other manuscript, this manuscript reads like an announcement of the challenge results rather than a research paper. The title of the manuscript says "machine learning approaches" but I had hard time to identify new machine learning approaches in the manuscript. What's more concerning is the lack of method description. All the methods from different teams were described at very high level without any details. It is difficult to evaluate.

- The number of submissions is also low (<10), raising the concern whether this is a successful challenge and how representative the methods are especially given that the performance seems to vary dramatically.

- The scope of the manuscript is narrow. It is unclear how generalizable the findings for performance evaluation in this manuscript are given that not every lab would produce intMEMOIR data. The authors should justify.

- The discrepancy between the results from intMEMOIR and the ground truth tree should be further exploited. The gap is coming from the computational methods or the intMEMOIR itself?

Reviewers' comments, manuscript 2:

Reviewer #1: This is a report of DREAM Challenge on cell lineage reconstruction based on cumulative Cas9 edits. It provides an overview on six different algorithms by DCLEAR(WHD), DCLEAR(KRD), Liu, Guan, Cassiopeia, and AMbeRland. Probably due to length limit, I found lack of sufficient info in the description of most algorithms. It is also disappointing to see none of the methods beat FastTree2 in the mouse challenge. Nonetheless, DCLEAR is impressive and better than FastTree2 in the C. elegans challenge.

Specific concerns:

1. Since DCLEAR is the winner, one should determine what types of errors DCLEAR often made. Such info may provide clues for future improvement.
2. It is unclear if most methods treat inter-target deletions as dropouts. If true, it is interesting to see Guan's method replacing all gap mutations with the mutation types. One should determine the benefits (if any) of including gap mutations for analysis.
3. Figure 4 (except D,F) is too general to be useful. One should consider using a same simplified input to illustrate these different algorithms with real numbers.
4. Figure 5B: Why are some nodes/edges labeled in red?
5. Figure 5C,D: Need more info in both illustrations and associated legends. Not sure what the single-sentence legend of 5D means.
6. Figure 6: A/B swapped.
7. Figure S3C: Both figure and legend are confusing.

Reviewer #2: In this manuscript Gong et al. reported the results from the Allen Institute lineage reconstruction DREAM challenge. The goal of the challenge is to reconstruct cell lineage trees based on phylogenetic approaches by considering the mutations induced by CRISPR. The results derived from the challenge with two different datasets were reported (C. elegans and a simulation of mouse development).

The goal of the challenge is to evaluate different tree-building approaches from the community. Needless to say, the overall effort was meaningful and the insights and methodologies from the challenges would be potentially useful for a wide research community. But as the manuscript currently stands, I have the following serious concerns of the work.

- First and foremost, this manuscript does not read like a research paper. It is more of an announcement of the challenge results. From the descriptions, almost all methods were previously known with little novelty. From the standpoint of rigor, the manuscript only describes the methods from different teams at high level (e.g., what metric was used what machine learning framework was employed) and does not have any detailed description of the methods submitted to the challenge. It is therefore impossible for the readers (and for me as a reviewer) to understand how all the methods were designed. The supplement simply lists all the methods in a table and GitHub links without any method description. In addition to much more detailed method description of all the approaches, the manuscript should carefully describe how these methods were run to generate the results to enhance reproducibility.

- The title says "machine learning approaches" but all the methods mentioned in the manuscript do not have new machine learning techniques. All are from existing approaches.

- Compared with earlier challenges with other problems, the number of teams seems quite low. It is therefore questionable whether the methods well represent the field and if the results derived from these methods are strong enough to reach meaningful conclusions. In the past, when the number of submissions was low, the organizers would implement some widely used approaches. This doesn't seem to be the case in this work (the organizers only wrote some evaluation code). Understanding how representative the methods are is further complicated by the fact the authors did not provide detailed method descriptions.

- The authors generated consensus trees for all methods. This is an interesting effort. I would suggest the authors make a major step further to offer generalizable recommended and/or consensus methods (rather than just for two datasets) for tree building for cell lineage trees for the community.

- Following the previous point, I would expect that any reader of this manuscript would be able to understand the following and also potentially use in their own work. The scope of the manuscript is too narrow at the moment.
1) How to generate simulated data and also use existing data to evaluate the methods by reproducing the results in the challenge? (this has been established in the manuscript although the work lacks details)
2) How to modify the simulation code to generate their own simulated datasets for different purpose (e.g., evaluating their own new development)? This is very important component to enhance the impact of the work in the community. Previous challenges in the genomics community such as genome assembly demonstrated how important this is.
3) What type of method should be used for different types of data and settings? How these methods actually work?
4) Is it possible to integrate different methods to generate consensus tree with better results? What are the considerations?
If the manuscript addresses these questions I believe it would be useful for the community. Right now, the

manuscript reads like a summary of the challenge without much useful information for the readers if they want to use them for their own projects.

Minor
- Many typos in the manuscript. Even in the abstract there are two obvious ones:
phylogenetical -> phylogenetic
challenfes -> challenges

Reviewer #3: This paper describes a recent DREAM challenge for reconstructing cell lineages. There are two challenges: lineage reconstruction for (1) 1,000 cells from C. elegans and (2) 10,000 cells from mouse. While the tree for the first challenge was based on a shuffled (via SPR moves), experimentally derived lineage tree of C elegans, the second challenge used an in silico tree generated following Stochastic Tree Grammar process. Both challenges consisted of in silico readouts of a 200 and 1000 CRISPR character array, accounting for varying mutation rates, single-cell drop out rates, and homozygous deletions due to CRISPR. There were 5 submissions for the first challenge and 3 submissions for the second challenge. My main concerns with this paper are (i) the data is almost fully in silico (unlike the other DREAM challenge by the same authors) and (ii) the number of submissions is fairly limited. Particularly for the second challenge, which considers a larger tree, there were only 3 submissions. Given that the field is headed towards large-scale lineage reconstruction, scalability of methods will be an issue. In addition, the writing at times is confusing and incomplete. Please see below for more details.

* Describe data more carefully.

Please provide a table with characteristics of the training and test data. It is not clear if sampled cells are stratified by time point. It is also not clear whether timestamp information is provided as input. It is not clear what type of edits occur in the CRISPR array: only deletions, or also insertions? Why are there 30 states for each character (A-Z, a-d)? What do these states correspond to? How long is each character in the genome (number of bases). Can a character change state multiple times?

* What is Fig 2A? Not referred to. Also, what are mutational outcomes?

* Abstract would benefit from some proof reading

* Title: "C elegans and mouse" => make consistent.

* How many cell division? 1-year in mouse, what about C elegans?

* Figure S1A: consense => consensus

* Neighbor Joining -- cite original paper: doi:10.1093/oxfordjournals.molbev.a040454

Cell Systems

Transparent Peer Review Record

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Pablo,

I hope this email finds you well. The reviews are back on your manuscript and I've appended them below. You'll see that this version of the paper is much better received than the previous versions, but there's still a bit of tightening up that needs to happen before we're ready to publish the paper. To help with this, I've made a few notes directly on the reviews and highlighted points that seem to warrant special attention. If you have any questions or concerns about this round of revision, which will be the last substantive one, I'd be happy to talk about them. I look forward to seeing your revised manuscript.

All the best,
Quincey

Quincey Justman, Ph.D.
Editor-in-Chief, *Cell Systems*

Reviewers' comments:

Reviewer #1: While the strength of the manuscript has been increased by merging the two previously separate manuscripts, I remain unconvinced about the generalizability of the challenge (i.e. intMEMOIR) and methods to current sequence-based data. I would appreciate more discussion about this in the manuscript itself. ***[From QJ: Please expand the Discussion to include these points.]*** In addition, I have two minor comments.

1. Non-uniqueness of barcodes and effect on RF and triplet distance. I do appreciate the authors new text in the manuscript about degeneracy and the new Fig. S3. However, I think that the RF distance (and triplet distance as well) can be adapted to consider splits in terms of barcodes/genotypes rather than cell labels. This requires looking at the symmetric different of multi-sets rather than sets. Please see supplement A.1.3 of the following paper.

[1] El-Kebir M, Satas G, Raphael BJ. Inferring parsimonious migration histories for metastatic cancers. Nat Genet. 2018;50(5):718-726. doi:10.1038/s41588-018-0106-z

Alternatively, one could think about solving an optimization problem to find a remapping of leaf/cell labels with identical barcodes so as to minimize RF distance. Given the high fraction of non-uniqueness as shown in Table S1, I encourage the authors to pursue either of these two options. I would like to see how well each method performs in light of what the authors call "theoretical maximum".

[From QJ: Please provide a fulsome response to Reviewer 1's point 1 in your point-by-point response, which will be freely available in the published paper's peer review record. Please use your judgement about incorporating these points into the main manuscript; I'd be happy to discuss and advise if you think this presents a tricky judgement call.]

2. Table of competing methods. It would also be good to briefly summarize the features/differences of the participating methods in a table.

[From QJ: In my view, this is optional.]

* Abstract: ground true trees => ground truth trees

Reviewer #2: This combined version of the manuscript to report the results of the DREAM challenge on cell lineage reconstructions is improved compared with the two separate earlier versions. I appreciate the authors' effort to rework the manuscript which, in my opinion, is now appropriate for publication after additional revision. I have the following comments for the authors to consider.

[From QJ: The points highlighted in yellow below are absolutely critical.]

1. As a manuscript reporting a challenge result and a resource, the organization of the resource, including code and datasets, and especially documentation was poorly performed. In the manuscript, there are numerous GitHub links to direct the readers to different GitHub repositories. Please do the following:

a) Create a centralized GitHub repository with a detailed Wiki page on GitHub to carefully describe the procedures to reproduce the results, generate data, and perform evaluations. This is currently scattered in various pages/links which is impossible for any interested reader to follow without monumental effort.

b) Carefully complete the documentations. For example, it should not be acceptable to have completely empty documentation for the code used in the Challenge: https://github.com/Sage-Bionetworks-Challenges/Allen-DREAM-Challenge

c) Make the datasets, including the intMEMOIR data, publicly accessible. I was not able to access the following page; therefore I am not sure if this is already available.
https://www.synapse.org/#!Synapse:syn20821809

2. I still think the description of methods used in the Challenge is inadequate. Also, the code for reproducing the challenge results from different teams is not all available. For example, I did not find the code for Guan et al.'s method and a few other methods. I may be confused because there are so many different links in the manuscript. If they are already available, please point out explicitly and, as I suggested in the previous comment, organize the code in a centralized place with carefully prepared documentation.

Authors' response to the reviewers' first round comments

Attached.

Reviewers' comments D-20-00473:

Reviewer #1: This paper describes a recent DREAM challenge regarding cell lineage reconstruction, benchmarking nine participating methods. The interesting aspect of this paper is the use of a recording technology (intMEMOIR) that enables one to reconstruct cell lineages and mutation status of cells in real time. This paper might be of interest to the readership of this journal if the following concerns are addressed.

We thank the reviewer for the interest in our study.

1. Relationship of experimental technology used in this challenge and lineage reconstruction tasks that occur in practice.

The authors constructed traced lineages and mutations of 106 experiments using the intMEMOIR system. Specifically, they started each experiment with a single cell with 10 loci that may mutate into one of two states: (1) deletion or (2) inversion. Mutations are irreversible (although this is only stated in Figure 3 caption).

We clarified the irreversibility of the edits in the text and Box 2.

It is not clear to me how this experimental setup relates to cell lineage reconstruction tasks that occur in practice. The first paragraph of the introduction mentions lineage reconstruction for understanding development. This needs to be expanded on in several ways. First, other important applications include cancer phylogeny reconstruction, somatic mosaiscm in the brain, clonal heterogeneity in T cells, etc. These must be described. Second, the input data must be described as well as limitations. For instance, data can differ in omic type (RNA, DNA, methylation, etc.), sequencing technology (single-cell vs bulk, long reads vs. short reads, or FISH), types of mutations (single-nucleotide variant, copy-number aberrations, structural variants), etc.

We included these points in the introduction.

I encourage the authors to describe how their experimental design fits with lineage reconstruction tasks that occur in practice. Specifically, I'd like to know:

* How applicable are developed methods to these general settings?
* Are they only applicable to this data?
* What are some key ideas/best practices that carry over.

intMEMOIR is new technology that is difficult to compare at the molecular level to the sequence-base approaches for lineage reconstruction as it also shows differences such as the absence of accidental deletions or *dropouts*. However, we think that this manuscript, complemented by the results from the *in silico* approaches, does have generalizable conclusions such as the necessity of having well calibrated mutation rates to avoid too little mutations but also array degenerations, the utility of having a training set of smaller trees to optimize lineage reconstruction methods both when choosing the features to determine sister-cell probabilities or distances and optimizing clustering approaches. It also allowed for a clear interpretation of the effect of the two different metrics with different tree sizes. We extended the discussion on the points raised by this reviewer.

2. Describe training and test data more carefully

My other main concern is that the description of the training and test data must be improved considerably. I spent a fair amount of time to understand key details that are not described in the main text.

We performed a thorough description in the text and Box 2.

* Figure 3 caption states that edits are irreversible. This needs to be stated in the main text. This is a key restriction.

We clarified the irreversibility of the edits in the text and Box2.

* Please add all 106 trees to supplement.

We included the trees to the supplement see Table S1.

* Please provide a table with key statistics of the input, describing the number of leaves, and the number of unique leaves (in terms of genotype) for each experiment.

We included the trees to the supplement and the key statistics see Table S1.

* Do you measure mutation status of intermediary cells (e.g. internal nodes in the tree)?

Mutation status is only measured by smFISH at the end of the timelapse imaging experiment, this was also added to Box2. This was indicated in the text and Box 2.

* Do you measure time of occurrence of each node (or leaf) in the tree?

Time of occurrence is measured but was not used for this challenge.

* What about the last 24 hours? Do new mutations occur? Is lineage only traced in the first 36 hours?
Mutations were induced for the first 36 hrs of growth (approximately 3 cell divisions) and cells were then allowed to grow with no further changes in the recording arrays for an additional 24 hrs. At this point the arrays for each cell in the colony were read using FISH. This has been added to the manuscript.

* How was the split in training and test performed? Randomly? Stratified in terms of tree size? Please explain.

The full dataset was split into a training set of 76 colonies and a test set of 30 colonies, such that the median distribution of tree sizes and the performance of an in-house Maximum-Likelihood reconstruction algorithm was similar between the two sets. This has been added to the manuscript.

* Did participants have access to the test data?

Participants only had access to the training sets, we clarified this in the text.

* Why is there variability in number of leaves?

Not all the cells survive nor divide at similar rates. This has been added to the manuscript.

* How does intMEMOIR build upon MEMOIR?

intMEMOIR using an integrase system has 3 different states (unmodified, inverted or deleted) for each character array when MEMOIR had only 2 states. For all details and thorough description of this new version of MEMOIR see *Chow et al*.

* How large are the altered regions (in terms of base pairs)?

As described in detail in Chow *et al*, each array is made of 10 memory units that contain a barcode flanked by a pair of inverted *attP* site on one end and *attB* site on the other composed of different central dinucleotides each. The total number of barcode nucleotides is 500bp (+/-2bp) in order to allow detection by the FISH probe, on each side, 8 nucleotides for the *attP* and *attB* sites that have to be added.

* Figure 2: define large vs. small (how many leaves).
A large tree has more than 8 cells and a small one has less or equal to 8. It has been indicated in the main text.

3. Ambiguity in cell labels?

My final main concern is about ambiguity in genotypes of leaves. Fig S3 suggests that there may be multiple leaves that have identical genotypes. I'd like to know whether their are genotypes that are common to more than two leaves. If this is the case then it is not fair to use RF or triplet distance on all leaves, as there will be exist multiple trees that are identical when removing this type of ambiguity. Please describe.

As seen in Table S1 the proportion of trees that have cells with identical arrays is significant, but Figure S3D illustrates that the teams'scores show no particular dependence on the overall degeneracy of a tree. Also, this degeneracy affects all methods equally and fairness is maintained, but it also probably explains why the scores plateaued for all methods. Finally, degeneracy is a typical problem encountered in the data normally used for reconstructing cell lineages and so it is important that this feature was included in this challenge.

Minor:

* Line 130: readout => read out
* Line 138: was splitted => was split

We corrected the indicated typos.

Reviewer #2: Granados et al. supplements the results described in the other manuscript to report outcome from the Allen Institute lineage reconstruction DREAM challenge using the

intMEMOIR recording data, a synthetic imaging-readable lineage tracing technology. I have the following major concerns of this work. I generally have the same serious concerns of this manuscript as the other one reporting this DREAM challenge outcome. I list some of them below.

- I strongly think this manuscript should be combined with the other one to present an overall report of this DREAM challenge.

Following this reviewer suggestion, we combined the manuscripts.

- Similar to the other manuscript, this manuscript reads like an announcement of the challenge results rather than a research paper. The title of the manuscript says "machine learning approaches" but I had hard time to identify new machine learning approaches in the manuscript. What's more concerning is the lack of method description. All the methods from different teams were described at very high level without any details. It is difficult to evaluate.

We changed the title to reflect the results and detailed the two most innovative methods (AMbeRland and DCLEAR), as the maximum parsimony method *Cassiopeia* has already been published. Also access to the code and implementation details for all methods is available in the provided tables S2 and S3. The goal of the challenge was to establish a machine learning setup to benchmark methods and generate novel approaches for linear reconstruction. Although most of the methods did not use an ML approach per se, notably AMBerRland did and generated the best approach. Also, as detailed in Figure 3, most teams took advantage one way or another of the training set and in the context of the *in silico* challenges we show that a smaller training set is indeed useful to calibrate methods for solving a larger tree.

- The number of submissions is also low (<10), raising the concern whether this is a successful challenge and how representative the methods are especially given that the performance seems to vary dramatically.
Challenge participation varies widely and depends of the size of the community. For us at DREAM the most important feature beyond a certain level of participation is the quality and innovation of the approaches. Clearly lineage reconstruction is a new field, but out of 9 submitted solutions, the top 4 methods reached with very different approaches a performance plateau (probably due to the cells'arrays degeneracy). We think that in that sense the challenge is successful and also as now a benchmark is available to compare further methods.

- The scope of the manuscript is narrow. It is unclear how generalizable the findings for performance evaluation in this manuscript are given that not every lab would produce intMEMOIR data. The authors should justify.

We agree with this reviewer that intMEMOIR is new technology that is difficult to compare at the molecular level to the sequence-base approaches for lineage reconstruction and shows differences such as the absence of non-wanted deletions or *dropouts*. However, we think that this manuscript, complemented by the results from the *in silico* approaches, does have generalizable conclusions such as the utility of having a training set of smaller trees to optimize lineage reconstruction methods both when choosing the features to determine sister-cell probabilities and optimizing clustering approaches. We extended the discussion on the point raised by this reviewer.

- The discrepancy between the results from intMEMOIR and the ground truth tree should be further exploited. The gap is coming from the computational methods or the intMEMOIR itself?

We extended the discussion on the point raised by this reviewer.

Reviewers' comments D-20-00472:

Reviewer #1: This is a report of DREAM Challenge on cell lineage reconstruction based on cumulative Cas9 edits. It provides an overview on six different algorithms by DCLEAR(WHD), DCLEAR(KRD), Liu, Guan, Cassiopeia, and AMbeRland. Probably due to length limit, I found lack of sufficient info in the description of most algorithms. It is also disappointing to see none of the methods beat FastTree2 in the mouse challenge. Nonetheless, DCLEAR is impressive and better than FastTree2 in the C. elegans challenge.

We thank the reviewer on appreciating the results of this study.

Specific concerns:

1. Since DCLEAR is the winner, one should determine what types of errors DCLEAR often made. Such info may provide clues for future improvement.

The results from both challenge 2 and 3 show that DCLEAR scored less favorably in Robinson Foulds distance, compared to the Triplet distance, suggesting that DCLEAR may have weakness of detecting major branching events in the early cell division stages. Both WHD and KRD in DCLEAR package rely on the rare mutations to estimate the cell distances. During early cell division stages, however, the rare mutations are significantly less likely to be present in the sequences and result in difficulties of separating early branching events. Modeling the dependence between multiple non-adjacent mutations in the sequences (on top of the neighboring *k*-mers) may be necessary to more accurately evaluate the early branching events. We added this to the discussion.

2. It is unclear if most methods treat inter-target deletions as dropouts. If true, it is interesting to see Guan's method replacing all gap mutations with the mutation types. One should determine the benefits (if any) of including gap mutations for analysis.

Other teams like *Liu Lab* just filled the gaps with the initial character 0, while in both *DCLEAR* WHD and KRD, the deletion and dropout were treated differently. In WHD, the weight for deletion, dropout, regular state and ground state are 0.9, 0.4, 3 and 1, respectively. In KRD, deletion and dropout are treated as two different characters. We added this to the discussion.

3. Figure 4 (except D,F) is too general to be useful. One should consider using a same simplified input to illustrate these different algorithms with real numbers.

We updated the figure and Fig.4D and it is now explained in Fig.5A, Fig.4F is now in Box2.

4. Figure 5B: Why are some nodes/edges labeled in red?

The red nodes represent an example referred in the text illustrating nodal distance, we updated the legend to clarify this.

5. Figure 5C,D: Need more info in both illustrations and associated legends. Not sure what the single-sentence legend of 5D means.

We extended the explanation for DCLEAR's method (see STAR methods) and updated the legend.

6. Figure 6: A/B swapped.

We took care of this omission now with the combined manuscripts.

7. Figure S3C: Both figure and legend are confusing.

We updated the figure and legend and also further explained AMbeRland's approach in the main text and Figure 5.

Reviewer #2: In this manuscript Gong et al. reported the results from the Allen Institute lineage reconstruction DREAM challenge. The goal of the challenge is to reconstruct cell lineage trees based on phylogenetic approaches by considering the mutations induced by CRISPR. The results derived from the challenge with two different datasets were reported (C. elegans and a simulation of mouse development). The goal of the challenge is to evaluate different tree-building approaches from the community. Needless to say, the overall effort was meaningful and the insights and methodologies from the challenges would be potentially useful for a wide research community. But as the manuscript currently stands, I have the following serious concerns of the work.

We thank the reviewer on appreciating the results of this study.

- First and foremost, this manuscript does not read like a research paper. It is more of an announcement of the challenge results. From the descriptions, almost all methods were previously known with little novelty. From the standpoint of rigor, the manuscript only describes the methods from different teams at high level (e.g., what metric was used what machine learning framework was employed) and does not have any detailed description of the methods submitted to the challenge. It is therefore impossible for the readers (and for me as a reviewer) to understand how all the methods were designed. The supplement simply lists all the methods in a table and GitHub links without any method description. In addition to much more detailed method description of all the approaches, the manuscript should carefully describe how these methods were run to generate the results to enhance reproducibility.

The main goal of the DREAM challenge is to describe and analyze the results obtained while benchmarking methods for lineage reconstruction. We also reproduced the results of every method using the provided code and described the most interesting/original characteristics of the

implementations. We further detailed in separate figures the non-published methods we considered the most interesting such as DCLEAR and AMbeRland (*Cassiopeia* although adapted for the challenge was previously published) as all details for other methods will be published separately. We hope the reviewer finds this is sufficient information.

- The title says "machine learning approaches" but all the methods mentioned in the manuscript do not have new machine learning techniques. All are from existing approaches.

To our knowledge this is the first attempt to evaluate lineage reconstruction algorithms using a machine learning scheme. Also, both DCLEAR and *AMberLand* are original methods and *Casssiopeia* was recently published and fully adapted for this challenge. Also, we are not aware of decision trees, the machine learning method implemented by *AMberLand*, having been applied for lineage reconstruction. Still, we changed the title to reflect this reviewer's comment.

- Compared with earlier challenges with other problems, the number of teams seems quite low. It is therefore questionable whether the methods well represent the field and if the results derived from these methods are strong enough to reach meaningful conclusions. In the past, when the number of submissions was low, the organizers would implement some widely used approaches. This doesn't seem to be the case in this work (the organizers only wrote some evaluation code). Understanding how representative the methods are is further complicated by the fact the authors did not provide detailed method descriptions.

Challenge participation varies widely and depends on the size of the community. For us at DREAM the most important feature beyond a certain level of participation is the quality and innovation of the approaches. Clearly lineage reconstruction is a new field, but in our view the 5 methods for the *C.elegans* challenge and 3 methods (+ 5 more for the 6000 cells tree of the leaderboard) for the much larger *M.musculus* tree were quite diverse (Maximum parsimony, two distance based methods and a decision-tree based method) and performed very well, although DCLEAR clearly outperformed all methods. Also we did provide the implementation of Fast2tree as comparison and constructed the consensus solutions, we also implemented but did not show NJ and TripleMaxCut as the results where worse than Fast2tree. We think that in that sense the challenge is successful and also as now a benchmark is available to compare further methods.

- The authors generated consensus trees for all methods. This is an interesting effort. I would suggest the authors make a major step further to offer generalizable recommended and/or consensus methods (rather than just for two datasets) for tree building for cell lineage trees for the community.

- Following the previous point, I would expect that any reader of this manuscript would be able to understand the following and also potentially use in their own work. The scope of the manuscript is too narrow at the moment.

1) How to generate simulated data and also use existing data to evaluate the methods by reproducing the results in the challenge? (this has been established in the manuscript although the work lacks details).

The simulated data was adapted from two publications Salvador *et al* for *C.elegans* and Spiro et al for *M.musculus* where all the details for reproducibility are available, furthermore the STAR methods indicate the main lines for reproducing the simulations and also point to the github sites where we also made available all the code with clear step by step indications for generating the datasets and also for scoring them as well as clear indications on how to run the participant's codes to reproduce their results.

2) How to modify the simulation code to generate their own simulated datasets for different purpose (e.g., evaluating their own new development)? This is very important component to enhance the impact of the work in the community. Previous challenges in the genomics community such as genome assembly demonstrated how important this is.

We consider that providing the code used to simulate the trees used in this challenge and the articles detailing their implementations, plus a detailed description on how we adapted those to generate the challenge *in silico* datasets should be enough for adapting the approach to any new problem (see STAR methods).

3) What type of method should be used for different types of data and settings? How these methods actually work?

We have now adapted the discussion to clarify that *DCLEAR* is the best adapted method for large trees and performs better for triplets than RF, but for smaller trees *AMberLand's* GBM approach with threshold selection is the best adapted. We also discussed how each metric explains how different algorithms perform, a better triplet performance pointing towards a better local structure in large trees as RF points towards earlier branching events.

4) Is it possible to integrate different methods to generate consensus tree with better results? What are the considerations?

We would have liked to develop a clear strategy for aggregating different types of methods, but we have no clear answer beyond the observation that the aggregation of all methods surprisingly generates very good results when measured with the RF distance but terrible when measured with triplet distances. Specifically, given that for example the maximum parsimony method scores very well in the triplets metric, while other distance based methods rank better with the RF metric, a combination of both could have brought the best results in both metrics.

If the manuscript addresses these questions I believe it would be useful for the community. Right now, the manuscript reads like a summary of the challenge without much useful information for the readers if they want to use them for their own projects.

Minor
- Many typos in the manuscript. Even in the abstract there are two obvious ones:
phylogenetical -> phylogenetic
challenfes -> challenges

We are sorry about these typos.

Reviewer #3: This paper describes a recent DREAM challenge for reconstructing cell lineages. There are two challenges: lineage reconstruction for (1) 1,000 cells from C. elegans and (2) 10,000 cells from mouse. While the tree for the first challenge was based on a shuffled (via SPR moves), experimentally derived lineage tree of C elegans, the second challenge used an in silico tree generated following Stochastic Tree Grammar process. Both challenges consisted of in silico readouts of a 200 and 1000 CRISPR character array, accounting for varying mutation rates, single-cell drop out rates, and homozygous deletions due to CRISPR. There were 5 submissions for the first challenge and 3 submissions for the second challenge. My main concerns with this paper are (i) the data is almost fully in silico (unlike the other DREAM challenge by the same authors) and (ii) the number of submissions is fairly limited. Particularly for the second challenge, which considers a larger tree, there were only 3 submissions. Given that the field is headed towards large-scale lineage reconstruction, scalability of methods will be an issue. In addition, the writing at times is confusing and incomplete. Please see below for more details.

The results of the *in silico* challenges have now been merged with the experimental-based challenge using *in vitro* grown cell lines. Also, challenge participation varies widely and depends on the size of the community. For us at DREAM the most important feature beyond a certain level of participation is the quality and innovation of the approaches. Clearly lineage reconstruction is a new field, but in our view the 5+1 methods for the *C.elegans* challenge and 3 methods for the much larger *M.musculus* tree (+ 5 submissions to the 6000 cell leaderboard) were quite diverse (Maximum parsimony, two distance based methods and a decision-tree based method) and performed very well, although DCLEAR clearly outperformed all methods. The *in vitro* challenge had 9 teams submitting, and there were 4 more submissions to the intermediary leaderboard for the *M.musculus* tree, totaling overall 22 submissions in the 3 challenges.

* Describe data more carefully.

We tried to describe the datasets with more clarity and specified it together with the definitions of the training/test sets in Box2.

Please provide a table with characteristics of the training and test data. It is not clear if sampled cells are stratified by time point. It is also not clear whether timestamp information is provided as input. It is not clear what type of edits occur in the CRISPR array: only deletions, or also insertions? Why are there 30 states for each character (A-Z, a-d)? What do these states correspond to? How long is each character in the genome (number of bases). Can a character change state multiple times?

We tried to describe the datasets with more clarity and specified it together with the definitions of the training/test sets in Box2. Specifically, only deletions and dropouts (for *M.musculus*) were modeled, the number of characters (30) was chosen in order to have sufficient diversity to reconstruct trees of 1000 and 10000 cells and characters are just proxies for bases in the genome and are not sets of nucleotides *per se* and character mutations is irreversible and happens only once.

* What is Fig 2A? Not referred to. Also, what are mutational outcomes?

We tried to better describe the figure now in Box 2. Mutational outcomes are the 30 possible irreversible mutations.

* Abstract would benefit from some proof reading

We apologize for typos, we tried this time to be more careful.

* Title: "C elegans and mouse" => make consistent.

We maintained consistency as well noted by this reviewer.

* How many cell division? 1-year in mouse, what about C elegans?

The *C.elegans* tree size and number of cell divisions is predetermined as it was solved experimentally, for the mouse it is also difficult to estimate given that not all cells are accounted for in the simulation (only about 10000 out of 10^{12} estimated).

* Figure S1A: consense => consensus

We corrected this typo.

* Neighbor Joining -- cite original paper: doi:10.1093/oxfordjournals.molbev.a040454

We thank the reviewer for indicating this citation.

Pablo Meyer, Ph.D.
Team Leader
IBM TJ Watson Research Center
1101 Kitchawan Road
Yorktown Heights, NY 10598
pmeyerr@us.ibm.com

Reviewers' comments D-20-00473:

Reviewer #1: This paper describes a recent DREAM challenge regarding cell lineage reconstruction, benchmarking nine participating methods. The interesting aspect of this paper is the use of a recording technology (intMEMOIR) that enables one to reconstruct cell lineages and mutation status of cells in real time. This paper might be of interest to the readership of this journal if the following concerns are addressed.

We thank the reviewer for the interest in our study.

1. Relationship of experimental technology used in this challenge and lineage reconstruction tasks that occur in practice.

The authors constructed traced lineages and mutations of 106 experiments using the intMEMOIR system. Specifically, they started each experiment with a single cell with 10 loci that may mutate into one of two states: (1) deletion or (2) inversion. Mutations are irreversible (although this is only stated in Figure 3 caption).

We clarified the irreversibility of the edits in the text and Box 2.

It is not clear to me how this experimental setup relates to cell lineage reconstruction tasks that occur in practice. The first paragraph of the introduction mentions lineage reconstruction for understanding development. This needs to be expanded on in several ways. First, other important applications include cancer phylogeny reconstruction, somatic mosaiscm in the brain, clonal heterogeneity in T cells, etc. These must be described. Second, the input data must be described as well as limitations. For instance, data can differ in omic type (RNA, DNA, methylation, etc.), sequencing technology (single-cell vs bulk, long reads vs. short reads, or FISH), types of mutations (single-nucleotide variant, copy-number aberrations, structural variants), etc.

We included these points in the introduction.

I encourage the authors to describe how their experimental design fits with lineage reconstruction tasks that occur in practice. Specifically, I'd like to know:

* How applicable are developed methods to these general settings?
* Are they only applicable to this data?
* What are some key ideas/best practices that carry over.

intMEMOIR is new technology that is difficult to compare at the molecular level to the sequence-base approaches for lineage reconstruction as it also shows differences such as the absence of accidental deletions or *dropouts*. However, we think that this manuscript, complemented by the results from the *in silico* approaches, does have generalizable conclusions such as the necessity of having well calibrated mutation rates to avoid too little mutations but

also array degenerations, the utility of having a training set of smaller trees to optimize lineage reconstruction methods both when choosing the features to determine sister-cell probabilities or distances and optimizing clustering approaches. It also allowed for a clear interpretation of the effect of the two different metrics with different tree sizes. We extended the discussion on the points raised by this reviewer.

2. Describe training and test data more carefully

My other main concern is that the description of the training and test data must be improved considerably. I spent a fair amount of time to understand key details that are not described in the main text.

We performed a thorough description in the text and Box 2.

* Figure 3 caption states that edits are irreversible. This needs to be stated in the main text. This is a key restriction.

We clarified the irreversibility of the edits in the text and Box2.

* Please add all 106 trees to supplement.

We included the trees to the supplement see Table S1.

* Please provide a table with key statistics of the input, describing the number of leaves, and the number of unique leaves (in terms of genotype) for each experiment.

We included the trees to the supplement and the key statistics see Table S1.

* Do you measure mutation status of intermediary cells (e.g. internal nodes in the tree)?

Mutation status is only measured by smFISH at the end of the timelapse imaging experiment, this was also added to Box2. This was indicated in the text and Box 2.

* Do you measure time of occurrence of each node (or leaf) in the tree?

Time of occurrence is measured but was not used for this challenge.

* What about the last 24 hours? Do new mutations occur? Is lineage only traced in the first 36 hours?
Mutations were induced for the first 36 hrs of growth (approximately 3 cell divisions) and cells were then allowed to grow with no further changes in the recording arrays for an additional 24 hrs. At this point the arrays for each cell in the colony were read using FISH. This has been added to the manuscript.

* How was the split in training and test performed? Randomly? Stratified in terms of tree size? Please explain.

The full dataset was split into a training set of 76 colonies and a test set of 30 colonies, such that the median distribution of tree sizes and the performance of an in-house Maximum-Likelihood

reconstruction algorithm was similar between the two sets. This has been added to the manuscript.

* Did participants have access to the test data?

Participants only had access to the training sets, we clarified this in the text.

* Why is there variability in number of leaves?

Not all the cells survive nor divide at similar rates. This has been added to the manuscript.

* How does intMEMOIR build upon MEMOIR?

intMEMOIR using an integrase system has 3 different states (unmodified, inverted or deleted) for each character array when MEMOIR had only 2 states. For all details and thorough description of this new version of MEMOIR see *Chow et al*.

* How large are the altered regions (in terms of base pairs)?

As described in detail in Chow *et al*, each array is made of 10 memory units that contain a barcode flanked by a pair of inverted *attP* site on one end and *attB* site on the other composed of different central dinucleotides each. The total number of barcode nucleotides is 500bp (+/-2bp) in order to allow detection by the FISH probe, on each side, 8 nucleotides for the *attP* and *attB* sites that have to be added.

* Figure 2: define large vs. small (how many leaves).
A large tree has more than 8 cells and a small one has less or equal to 8. It has been indicated in the main text.

3. Ambiguity in cell labels?

My final main concern is about ambiguity in genotypes of leaves. Fig S3 suggests that there may be multiple leaves that have identical genotypes. I'd like to know whether their are genotypes that are common to more than two leaves. If this is the case then it is not fair to use RF or triplet distance on all leaves, as there will be exist multiple trees that are identical when removing this type of ambiguity. Please describe.

As seen in Table S1 the proportion of trees that have cells with identical arrays is significant, but Figure S3D illustrates that the teams'scores show no particular dependence on the overall degeneracy of a tree. Also, this degeneracy affects all methods equally and fairness is maintained, but it also probably explains why the scores plateaued for all methods. Finally, degeneracy is a typical problem encountered in the data normally used for reconstructing cell lineages and so it is important that this feature was included in this challenge.

Minor:

* Line 130: readout => read out
* Line 138: was splitted => was split

We corrected the indicated typos.

Reviewer #2: Granados et al. supplements the results described in the other manuscript to report outcome from the Allen Institute lineage reconstruction DREAM challenge using the intMEMOIR recording data, a synthetic imaging-readable lineage tracing technology. I have the following major concerns of this work. I generally have the same serious concerns of this manuscript as the other one reporting this DREAM challenge outcome. I list some of them below.

- I strongly think this manuscript should be combined with the other one to present an overall report of this DREAM challenge.

Following this reviewer suggestion, we combined the manuscripts.

- Similar to the other manuscript, this manuscript reads like an announcement of the challenge results rather than a research paper. The title of the manuscript says "machine learning approaches" but I had hard time to identify new machine learning approaches in the manuscript. What's more concerning is the lack of method description. All the methods from different teams were described at very high level without any details. It is difficult to evaluate.

We changed the title to reflect the results and detailed the two most innovative methods (AMbeRland and DCLEAR), as the maximum parsimony method *Cassiopeia* has already been published. Also access to the code and implementation details for all methods is available in the provided tables S2 and S3. The goal of the challenge was to establish a machine learning setup to benchmark methods and generate novel approaches for linear reconstruction. Although most of the methods did not use an ML approach per se, notably AMBerRland did and generated the best approach. Also, as detailed in Figure 3, most teams took advantage one way or another of the training set and in the context of the *in silico* challenges we show that a smaller training set is indeed useful to calibrate methods for solving a larger tree.

- The number of submissions is also low (<10), raising the concern whether this is a successful challenge and how representative the methods are especially given that the performance seems to vary dramatically.
Challenge participation varies widely and depends of the size of the community. For us at DREAM the most important feature beyond a certain level of participation is the quality and innovation of the approaches. Clearly lineage reconstruction is a new field, but out of 9 submitted solutions, the top 4 methods reached with very different approaches a performance plateau (probably due to the cells'arrays degeneracy). We think that in that sense the challenge is successful and also as now a benchmark is available to compare further methods.

- The scope of the manuscript is narrow. It is unclear how generalizable the findings for performance evaluation in this manuscript are given that not every lab would produce intMEMOIR data. The authors should justify.

We agree with this reviewer that intMEMOIR is new technology that is difficult to compare at the molecular level to the sequence-base approaches for lineage reconstruction and shows differences such as the absence of non-wanted deletions or *dropouts*. However, we think that this manuscript, complemented by the results from the *in silico* approaches, does have generalizable conclusions such as the utility of having a training set of smaller trees to optimize lineage reconstruction methods both when choosing the features to determine sister-cell probabilities and

optimizing clustering approaches. We extended the discussion on the point raised by this reviewer.

- The discrepancy between the results from intMEMOIR and the ground truth tree should be further exploited. The gap is coming from the computational methods or the intMEMOIR itself?

We extended the discussion on the point raised by this reviewer.

Reviewers' comments **D-20-00472**:
Reviewer #1: This is a report of DREAM Challenge on cell lineage reconstruction based on cumulative Cas9 edits. It provides an overview on six different algorithms by DCLEAR(WHD), DCLEAR(KRD), Liu, Guan, Cassiopeia, and AMbeRland. Probably due to length limit, I found lack of sufficient info in the description of most algorithms. It is also disappointing to see none of the methods beat FastTree2 in the mouse challenge. Nonetheless, DCLEAR is impressive and better than FastTree2 in the C. elegans challenge.

We thank the reviewer on appreciating the results of this study.

Specific concerns:

1. Since DCLEAR is the winner, one should determine what types of errors DCLEAR often made. Such info may provide clues for future improvement.

The results from both challenge 2 and 3 show that DCLEAR scored less favorably in Robinson Foulds distance, compared to the Triplet distance, suggesting that DCLEAR may have weakness of detecting major branching events in the early cell division stages. Both WHD and KRD in DCLEAR package rely on the rare mutations to estimate the cell distances. During early cell division stages, however, the rare mutations are significantly less likely to be present in the sequences and result in difficulties of separating early branching events. Modeling the dependence between multiple non-adjacent mutations in the sequences (on top of the neighboring *k*-mers) may be necessary to more accurately evaluate the early branching events. We added this to the discussion.

2. It is unclear if most methods treat inter-target deletions as dropouts. If true, it is interesting to see Guan's method replacing all gap mutations with the mutation types. One should determine the benefits (if any) of including gap mutations for analysis.

Other teams like *Liu Lab* just filled the gaps with the initial character 0, while in both *DCLEAR* WHD and KRD, the deletion and dropout were treated differently. In WHD, the weight for deletion, dropout, regular state and ground state are 0.9, 0.4, 3 and 1, respectively. In KRD, deletion and dropout are treated as two different characters. We added this to the discussion.

3. Figure 4 (except D,F) is too general to be useful. One should consider using a same simplified input to illustrate these different algorithms with real numbers.

We updated the figure and Fig.4D and it is now explained in Fig.5A, Fig.4F is now in Box2.

4. Figure 5B: Why are some nodes/edges labeled in red?

The red nodes represent an example referred in the text illustrating nodal distance, we updated the legend to clarify this.

5. Figure 5C,D: Need more info in both illustrations and associated legends. Not sure what the single-sentence legend of 5D means.

We extended the explanation for DCLEAR's method (see STAR methods) and updated the legend.

6. Figure 6: A/B swapped.

We took care of this omission now with the combined manuscripts.

7. Figure S3C: Both figure and legend are confusing.

We updated the figure and legend and also further explained AMbeRland's approach in the main text and Figure 5.

Reviewer #2: In this manuscript Gong et al. reported the results from the Allen Institute lineage reconstruction DREAM challenge. The goal of the challenge is to reconstruct cell lineage trees based on phylogenetic approaches by considering the mutations induced by CRISPR. The results derived from the challenge with two different datasets were reported (C. elegans and a simulation of mouse development). The goal of the challenge is to evaluate different tree-building approaches from the community. Needless to say, the overall effort was meaningful and the insights and methodologies from the challenges would be potentially useful for a wide research community. But as the manuscript currently stands, I have the following serious concerns of the work.

We thank the reviewer on appreciating the results of this study.

- First and foremost, this manuscript does not read like a research paper. It is more of an announcement of the challenge results. From the descriptions, almost all methods were previously known with little novelty. From the standpoint of rigor, the manuscript only describes the methods from different teams at high level (e.g., what metric was used what machine learning framework was employed) and does not have any detailed description of the methods submitted to the challenge. It is therefore impossible for the readers (and for me as a reviewer) to understand how all the methods were designed. The supplement simply lists all the methods in a table and GitHub links without any method description. In addition to much more detailed method description of all the approaches, the manuscript should carefully describe how these methods were run to generate the results to enhance reproducibility.

The main goal of the DREAM challenge is to describe and analyze the results obtained while benchmarking methods for lineage reconstruction. We also reproduced the results of every method using the provided code and described the most interesting/original characteristics of the implementations. We further detailed in separate figures the non-published methods we considered the most interesting such as DCLEAR and AMbeRland (*Cassiopeia* although adapted for the challenge was previously published) as all details for other methods will be published separately. We hope the reviewer finds this is sufficient information.

- The title says "machine learning approaches" but all the methods mentioned in the manuscript do not have new machine learning techniques. All are from existing approaches.

To our knowledge this is the first attempt to evaluate lineage reconstruction algorithms using a machine learning scheme. Also, both DCLEAR and *AMberLand* are original methods and *Casssiopeia* was recently published and fully adapted for this challenge. Also, we are not aware of decision trees, the machine learning method implemented by *AMberLand,* having been applied for lineage reconstruction. Still, we changed the title to reflect this reviewer's comment.

- Compared with earlier challenges with other problems, the number of teams seems quite low. It is therefore questionable whether the methods well represent the field and if the results derived from these methods are strong enough to reach meaningful conclusions. In the past, when the number of submissions was low, the organizers would implement some widely used approaches. This doesn't seem to be the case in this work (the organizers only wrote some evaluation code). Understanding how representative the methods are is further complicated by the fact the authors did not provide detailed method descriptions.

Challenge participation varies widely and depends on the size of the community. For us at DREAM the most important feature beyond a certain level of participation is the quality and innovation of the approaches. Clearly lineage reconstruction is a new field, but in our view the 5 methods for the *C.elegans* challenge and 3 methods (+ 5 more for the 6000 cells tree of the leaderboard) for the much larger *M.musculus* tree were quite diverse (Maximum parsimony, two distance based methods and a decision-tree based method) and performed very well, although DCLEAR clearly outperformed all methods. Also we did provide the implementation of Fast2tree as comparison and constructed the consensus solutions, we also implemented but did not show NJ and TripleMaxCut as the results where worse than Fast2tree. We think that in that sense the challenge is successful and also as now a benchmark is available to compare further methods.

- The authors generated consensus trees for all methods. This is an interesting effort. I would suggest the authors make a major step further to offer generalizable recommended and/or consensus methods (rather than just for two datasets) for tree building for cell lineage trees for the community.

- Following the previous point, I would expect that any reader of this manuscript would be able to understand the following and also potentially use in their own work. The scope of the manuscript is too narrow at the moment.

1) How to generate simulated data and also use existing data to evaluate the methods by reproducing the results in the challenge? (this has been established in the manuscript although the work lacks details).

The simulated data was adapted from two publications Salvador *et al* for *C.elegans* and Spiro et al for *M.musculus* where all the details for reproducibility are available, furthermore the STAR methods indicate the main lines for reproducing the simulations and also point to the github sites where we also made available all the code with clear step by step indications for generating the datasets and also for scoring them as well as clear indications on how to run the participant's codes to reproduce their results.

2) How to modify the simulation code to generate their own simulated datasets for different purpose (e.g., evaluating their own new development)? This is very important component to enhance the impact of the work in the community. Previous challenges in the genomics community such as genome assembly demonstrated how important this is.

We consider that providing the code used to simulate the trees used in this challenge and the articles detailing their implementations, plus a detailed description on how we adapted those to generate the challenge *in silico* datasets should be enough for adapting the approach to any new problem (see STAR methods).

3) What type of method should be used for different types of data and settings? How these methods actually work?

We have now adapted the discussion to clarify that *DCLEAR* is the best adapted method for large trees and performs better for triplets than RF, but for smaller trees *AMberLand's* GBM approach with threshold selection is the best adapted. We also discussed how each metric explains how different algorithms perform, a better triplet performance pointing towards a better local structure in large trees as RF points towards earlier branching events.

4) Is it possible to integrate different methods to generate consensus tree with better results? What are the considerations?

We would have liked to develop a clear strategy for aggregating different types of methods, but we have no clear answer beyond the observation that the aggregation of all methods surprisingly generates very good results when measured with the RF distance but terrible when measured with triplet distances. Specifically, given that for example the maximum parsimony method scores very well in the triplets metric, while other distance based methods rank better with the RF metric, a combination of both could have brought the best results in both metrics.

If the manuscript addresses these questions I believe it would be useful for the community. Right now, the manuscript reads like a summary of the challenge without much useful information for the readers if they want to use them for their own projects.

Minor
- Many typos in the manuscript. Even in the abstract there are two obvious ones:
phylogenetical -> phylogenetic
challenfes -> challenges

We are sorry about these typos.

Reviewer #3: This paper describes a recent DREAM challenge for reconstructing cell lineages. There are two challenges: lineage reconstruction for (1) 1,000 cells from C. elegans and (2) 10,000 cells from mouse. While the tree for the first challenge was based on a shuffled (via SPR moves), experimentally derived lineage tree of C elegans, the second challenge used an in silico tree generated following Stochastic Tree Grammar process. Both challenges consisted of in silico readouts of a 200 and 1000 CRISPR character array, accounting for varying mutation rates, single-cell drop out rates, and homozygous deletions due to CRISPR. There were 5 submissions for the first challenge and 3 submissions for the second challenge. My main concerns with this paper are (i) the data is almost fully in silico (unlike the other DREAM challenge by the same authors) and (ii) the number of submissions is fairly limited. Particularly for the second challenge, which considers a larger tree, there were only 3 submissions. Given that the field is headed towards large-scale lineage reconstruction, scalability of methods will be an issue. In addition, the writing at times is confusing and incomplete. Please see below for more details.

The results of the *in silico* challenges have now been merged with the experimental-based challenge using *in vitro* grown cell lines. Also, challenge participation varies widely and depends on the size of the community. For us at DREAM the most important feature beyond a certain level of participation is the quality and innovation of the approaches. Clearly lineage reconstruction is a new field, but in our view the 5+1 methods for the *C.elegans* challenge and 3 methods for the much larger *M.musculus* tree (+ 5 submissions to the 6000 cell leaderboard) were quite diverse (Maximum parsimony, two distance based methods and a decision-tree based method) and performed very well, although DCLEAR clearly outperformed all methods. The *in vitro* challenge had 9 teams submitting, and there were 4 more

submissions to the intermediary leaderboard for the *M.musculus* tree, totaling overall 22 submissions in the 3 challenges.

* Describe data more carefully.

We tried to describe the datasets with more clarity and specified it together with the definitions of the training/test sets in Box2.

Please provide a table with characteristics of the training and test data. It is not clear if sampled cells are stratified by time point. It is also not clear whether timestamp information is provided as input. It is not clear what type of edits occur in the CRISPR array: only deletions, or also insertions? Why are there 30 states for each character (A-Z, a-d)? What do these states correspond to? How long is each character in the genome (number of bases). Can a character change state multiple times?

We tried to describe the datasets with more clarity and specified it together with the definitions of the training/test sets in Box2. Specifically, only deletions and dropouts (for *M.musculus*) were modeled, the number of characters (30) was chosen in order to have sufficient diversity to reconstruct trees of 1000 and 10000 cells and characters are just proxies for bases in the genome and are not sets of nucleotides *per se* and character mutations is irreversible and happens only once.

* What is Fig 2A? Not referred to. Also, what are mutational outcomes?

We tried to better describe the figure now in Box 2. Mutational outcomes are the 30 possible irreversible mutations.

* Abstract would benefit from some proof reading

We apologize for typos, we tried this time to be more careful.

* Title: "C elegans and mouse" => make consistent.

We maintained consistency as well noted by this reviewer.

* How many cell division? 1-year in mouse, what about C elegans?

The *C.elegans* tree size and number of cell divisions is predetermined as it was solved experimentally, for the mouse it is also difficult to estimate given that not all cells are accounted for in the simulation (only about 10000 out of 10^{12} estimated).

* Figure S1A: consense => consensus

We corrected this typo.

* Neighbor Joining -- cite original paper: doi:10.1093/oxfordjournals.molbev.a040454

We thank the reviewer for indicating this citation.

Reviewers' comments D-20-00473R1:

Reviewer #1: While the strength of the manuscript has been increased by merging the two previously separate manuscripts, I remain unconvinced about the generalizability of the challenge (i.e. intMEMOIR) and methods to current sequence-based data. I would appreciate more discussion about this in the manuscript itself. ***[From QJ: Please expand the Discussion to include these points.]***

We thank the reviewer for appreciating the work we did in this revision and added the following paragraph to the discussion in response to his comment:

Regarding the generalization of the results obtained with the intMEMOIR technology which is difficult to compare at the molecular level to the sequence-based approaches for lineage reconstruction as it also shows differences such as the absence of accidental deletions or dropouts, we think that in conjunction with the results from the in silico approaches, the generalizable conclusions are the necessity of having well calibrated mutation rates to avoid too little mutations but also array degenerations, the utility of having a training set of smaller trees to optimize lineage reconstruction methods, distances and clustering, and allowing for a clear interpretation of the effect of the two different metrics with different tree sizes.

In addition, I have two minor comments.

1. Non-uniqueness of barcodes and effect on RF and triplet distance. I do appreciate the authors new text in the manuscript about degeneracy and the new Fig. S3. However, I think that the RF distance (and triplet distance as well) can be adapted to consider splits in terms of barcodes/genotypes rather than cell labels. This requires looking at the symmetric different of multi-sets rather than sets. Please see supplement A.1.3 of the following paper.

[1] El-Kebir M, Satas G, Raphael BJ. Inferring parsimonious migration histories for metastatic cancers. Nat Genet. 2018;50(5):718-726. doi:10.1038/s41588-018-0106-z

Alternatively, one could think about solving an optimization problem to find a remapping of leaf/cell labels with identical barcodes so as to minimize RF distance. Given the high fraction of non-uniqueness as shown in Table S1, I encourage the authors to pursue either of these two options. I would like to see how well each method performs in light of what the authors call "theoretical maximum".

We thank the reviewer for pushing the analysis regarding the impact of barcode degeneracy and in particular our interpretation of the plateauing results being a "theoretical maximum" due to the barcode degeneracy. In order to understand the effect in the predictions of such degeneracy, we implemented a simpler approach with a 100x bootstrap to merge all cells with identical barcodes into a single leaf (choosing each time a different cell with the same barcode) and recalculated the scores for each fold, all trees in the test set, and all best-performing teams. The comparison of team performance clearly shows an intially suprising result (see added panel Figure S3E) as merging cells with degenerate barcodes lowers the performance of all teams as measured by both metrics. This shows that the "theoretical maximum" does not come from the degeneracy but from the tree structure. *A posteriori* this makes sense because participants considered that degenerate

barcodes were sister cells and are thus most often part of the same partition generating a better RF distance. We modified the text accordingly as shown below:

*We thought this could be mainly due to the high degeneracy in cell arrays where two or more cells show identical edit patterns, but further analysis showed that barcode degeneration did not affect the performance of the teams (**Fig S3E**).*

2. Table of competing methods. It would also be good to briefly summarize the features/differences of the participating methods in a table.
Tables S2 and S3 contain this information.

* Abstract: ground true trees => ground truth trees

We made this change.

Reviewer #2: This combined version of the manuscript to report the results of the DREAM challenge on cell lineage reconstructions is improved compared with the two separate earlier versions. I appreciate the authors' effort to rework the manuscript which, in my opinion, is now appropriate for publication after additional revision. I have the following comments for the authors to consider.

1. As a manuscript reporting a challenge result and a resource, the organization of the resource, including code and datasets, and especially documentation was poorly performed. In the manuscript, there are numerous GitHub links to direct the readers to different GitHub repositories. Please do the following:
a) Create a centralized GitHub repository with a detailed Wiki page on GitHub to carefully describe the procedures to reproduce the results, generate data, and perform evaluations. This is currently scattered in various pages/links which is impossible for any interested reader to follow without monumental effort.
b) Carefully complete the documentations. For example, it should not be acceptable to have completely empty documentation for the code used in the Challenge:
https://github.com/Sage-Bionetworks-Challenges/Allen-DREAM-Challenge
c) Make the datasets, including the intMEMOIR data, publicly accessible. I was not able to access the following page; therefore I am not sure if this is already available.
https://www.synapse.org/#!Synapse:syn20821809

We were careful to follow the rightful indications made by this reviewer and now created a centralized wiki and completed all missing parts in the documentation. Also all as indicated in the wiki, the data is publicly accessible via signing-up to synapse.

2. I still think the description of methods used in the Challenge is inadequate. Also, the code for reproducing the challenge results from different teams is not all available. For example, I did not find the code for Guan et al.'s method and a few other methods. I may be confused because there are so many different links in the manuscript. If they are already available, please point out explicitly and, as I suggested in the previous comment, organize the code in a centralized place with carefully prepared documentation.
This paper describes a recent DREAM challenge regarding cell lineage reconstruction, benchmarking nine participating methods. The interesting aspect of this paper is the use of a recording technology (intMEMOIR) that enables one to reconstruct cell lineages and mutation status of cells in real time. This paper might be of interest to the readership of this journal if the

following concerns are addressed.

We thank the reviewer for the interest in our study and hope his important concerns regarding the establishing easier reproducibility are now adressed. We compiled all best performing the challenge related methods, data generation and scoring in a well documented wiki: https://github.com/Lineage-Reconstruction-DREAM-Challenge/hub/wiki

We hope the reviewer finds the effort adequate.