Response to Reviewers

March 13, 2023

1 Editor's Comments

Comment: *We are all satisfied with this new version, except for a few minor points that I encourage you to address before final recommendation. This should not take you too much time as a whole. Please address in particular the first point raised by the reviewer regarding SCARLET and either benchmark this method or argue better in favor of not doing it.*
Response: We appreciate the reviewer's observation about the reference we gave for the SA501 dataset (Eirew et al., 2015). While the model system, i.e., SA501 PDX line, was established in Eirew et al., 2015, the readout from the sequencing assay that we use in our manuscript is different, and was reported in Laks et al., 2019. Eirew et al., 2015 uses targeted DNA sequencing appropriate for calling point mutations. However, we use direct library preparation that is suitable for detecting copy number states. We apologize for this confusion and have corrected the reference in the manuscript. We further point out that SCARLET takes as one of its inputs, a phylogenetic tree infered from copy number data (c.f., Satas et al., 2020, Figure 2E). While such a tree is the output of our method. That is why a meaningful benchmark on DLP data for SCARLET is not feasible. Please also see our response to the reviewer comment.

Comment: *As far as I am concerned, I sincerely acknowledge your efforts in expanding explanations (best possible tree, proxy of violation rate + a wealth of terms and phrases) and adding new analyses: comparison with new methods and assessment of within-site pairwise dependencies - I appreciated your idea of getting rid of one of the two extremities of a segment that was gained or lost. However, here and in various other tests (violation of perfect phylogeny assumption, violation of infinite-site assumption), you seem to be reluctant to simulate the real biological process of CNA, as I had suggested in my report. Can you please explain me why?*
Response: Simulating realistic biological processes of CNA is an open research problem as it requires capturing various biological phenomena (DNA repair deficiency, chromosome missegregation, etc [1, 2]). We have, however, used a CNA simulator (Mallory, Xian F., et al. 2020) in our synthetic data experiments to benchmark against competing phylogenetic tree inference methods. This method simulates CN gains and losses along a tree. Notably, this

simulator allows daughter cells to inherit the CNs of their parent cell, and simulates extra events over this background. See Supplemental Figure 7 and 8 for simulated data and benchmark results respectively. We employed a second model (coalescent model; Section 9.5.3) for synthetic data experiments pertaining to violations of assumptions. We believe this is a valuable decision, as it diversifies our experiment settings under different models of data to cover a variety of scenarios in the absence of a ground truth.

[1] https://www.nature.com/articles/s41586-022-04789-9/figures/2
[2] https://www.nature.com/articles/s41586-022-04738-6

Comment: *Last point: line 610, it sounds a little weird to speak of the "three noise regimes" before explaining (in section 9.5.3) what they are.*

Response: We rephrased the sentence: "We performed the experiment described above on the S90 datasets (described in 9.5.3) with three noise regimes described as follows: . . . "

2 Reviewer's Comments

My questions and comments have largely been addressed by the authors. There are a few remaining comments I would like to make in response:

Comment: *1. In their response to my comment 5, in which I suggested to include SCARLET in the benchmark of methods, the authors wrote "Rationale of choice of additional baselines: sitka is designed for shallow sequencing regimes where calling SNVs per cell would be difficult, but copy numbers can be called reliably. In such cases, most SNVs will not be called in most cells. However SCARLET, while correcting for CNAs, requires the same SNV to be called in all cells."*

I am a bit surprised by this, since the SCARLET method explicitly accounts for allele drop-outs, i.e. missing SNV calls in some cells (cited from Satas et al., 2020): "Data from scDNA-seq typically have high error rates in identifying SNVs, and particularly high rates of false negatives and missing data due to amplification bias and allele dropout (Gawad et al., 2016). SCARLET models these errors using a beta-binominal distribution (Singer et al., 2018) of the observed read counts."

Further, in the SCARLET paper, the model was applied to a dataset from Leung et al. (2017), which has the following properties (cited from Satas et al., 2020):

"This data-set included targeted sequencing of 1,000 genes in 141 cells from a primary colon tumor and 45 cells from a matched liver metastasis (Figure 4A). The authors identified 36 SNVs and used SCITE (Jahn et al., 2016) to derive a perfect phylogeny from these SNVs (Figure 4B)."

These properties sounds very comparable to one of the cohorts, Eirew et al. (2015), that were analyzed in the current study:"DNA was prepared from 90 individual SA501 xenograft nuclei from passages X1, X2 and X4, and the

variant allele ratios were determined by targeted ultra-deep sequencing at 45 somatic SNV and 10 germline SNV positions."

Hence, I somewhat fail to see why the comparison to SCARLET was not even tried.

Response: This is an astute observation by the reviewer. In introducing the datasets in the manuscript, we had referenced the paper Eirew et al., 2015 where the biological substrate, i.e., the PDX line for SA501 was established, but crucially not the single cell DNA sequencing assay, namely direct library preparation (DLP). The reviewer correctly points out that in Eirew et al. (2015) targeted DNA sequencing is used, appropriate for calling point mutations. However, we use DLP readouts in our manuscript. DLP and similar *whole-genome* shallow single cell sequencing assays are the intended target of our method (c.f., Introduction, lines 56-63). We have clarified this in the manuscript, and cited Laks et al., 2019 where the results from applying DLP to SA501 and the OVA datasets used in our manuscript is described. We would like to also point out that SCARLET takes as input a tree generated from copy number profiles (Satas et al., 2020, Figure 2. panel E). However, such a tree is the output of our method. This, combined with the fact that SCARLET is designed for a different type of data (targeted DNA sequencing, appropriate for calling point mutations), we submit that benchmarking SCARLET on single cell copy number data is not feasible.

Comment: *2. I did not find a reference for the OVA dataset, even though it is not stated that that dataset was specifically generated for this study (as with the SA535 dataset).*

Response: We apologize for this omission and have added the following reference in text: E. Laks et al. "Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing". In: Cell 179.5 (2019), 1207–1221.e22.

Comment: *3. The authors further argued that the differences in performance between methods might be driven in part by the fact that some algorithms do not converge in the "available computational budget (several days)." I think it would be necessary to define the criteria for the allowed algorithm runtime/computational budget very clearly if runtime is such a crucial factor for performance. In other words, even though of course algorithm runtime is an important practical feature, the benchmark is supposed to measure accuracy, ideally after convergence and independent of runtime. If this cannot be achieved, it should be pointed out that the other methods may have achieved higher accuracy with a longer runtime.*

Response: We have amended section 2.2. to reflect the reviewer's comment as follows:

While due to limitation to our available computational budget, we could not allocate more time to the benchmarked methods, it is possible that given more runtime/computation budget, the other methods might have converged to more accurate solutions.

Comment: *4. Regarding my comment 6, I apologize for having failed to formulate the question in a manner that would have allowed the authors to understand and answer it properly. Even though the figure caption may have been a bit spartan, I would argue it was reasonably obvious that the previous Fig. 1f showed the insets from panel 1e. My question about what is now Fig. 3f regarded the (extinct) leaves without cells (i.e. those that do not have blue circles at the end). There are two such leaves in the box denoting iteration 100 and one in the box of iteration 101. Are these the unseen wild-type states of the marker events? Why are there two such extinct lines associated with just one marker (chr12_1600) in the left box? The process that happens in the upper part of the plot exactly corresponds to the description of the edge insertion process in the Methods section. However, it is not clear why in addition both the topology of the tree in the bottom part (presumably unaffected by the edge insertion) and the marker ID itself are changed (assuming each marker/red diamond is associated with its nearest orange text descriptor).*

Response: We thank the reviewer for pointing out this error in our visualization. In Figure 3, panel f, some of the leaf and marker labels were cut off. Moreover, an irrelevant marker name was displayed. We have fixed this and updated this panel to (1) properly show the cell and marker labels, and (2) to only show markers and leaf labels that are relevant to the edge insertion event.