Response to Reviewers

November 11, 2022

Thank you for your comments and for securing reviewers who provided valuable constructive criticisms. Below are our point-by-point responses to your comments as well as a summary of our responses to the referees' key points. The comments prompted us to make numerous improvements to the writing, perform new comparisons with additional recently developed methods, new experiments comparing MLE and Bayesian methods, as well as new experiments investigating the effect of ignoring independence across sites. All received comments are copied in this letter for convenience.

1 Recommender's comments

Comment: Recommender's summary.

Let me first give my own, self-contained summary of the manuscript. This paper presents and applies a new Bayesian inference method of phylogenetic reconstruction for multiple sequence alignments in the case of low sequencing coverage but diverse copy number aberrations (CNA), with applications to single cell sequencing of tumors. The idea is to take advantage of CNA to reconstruct the topology of the phylogenetic tree of sequenced cells in a first step (the 'sitka' method), and in a second step to assign single nucleotide variants (SNV) to tree edges (and then calibrate their lengths) (the 'sitka-snv' method). The data are assumed to be in the form of an integer-valued $C \times L$ matrix A, where C is the number of cells, L is the number of loci (here, loci are segments of prescribed length called 'bins'), whose entry a_{ij} at row i and column j is the (preprocessed) number of copies, called 'copy number state' (CNS), of locus j in cell i. These data are then summarized/simplified into a binary-valued $C \times L$ matrix Y, whose entry at row i and column j is $y_{ij} = 0$ if in cell i, the CN S at locus j and at locus $j + 1$ are equal (i.e., $a_{ij} = a_{i,j+1}$), and $y_{ij} = 1$ otherwise. Biologically speaking, when $y_{ij} = 1$, in the ancestral lineage of cell i, at least one genomic rearrangement has occurred, and more specifically the gain or loss of a segment with at least one endpoint in locus j or in locus $j + 1$; this event is viewed as a 'mutation at marker j', where marker j is the point where loci j and $j + 1$ touch. The authors expect the infinite-allele assumption to approximately hold (i.e., that at most one mutation occurs at any given marker 1 and that 0 is the ancestral state). They refer to this assumption *as the 'perfect phylogeny assumption'. By only recording from CNA events the endpoints at which they occur, the authors lose the information on copy number state (and also forget the dependencies between these endpoints), but they gain the assumption of independence of the mutational processes occurring at different sites, which approximately holds for CNA endpoints but certainly not for CNS. The goal of sitka is to produce a posterior distribution on phylogenetic trees conditional on the matrix Y, where here a phylogenetic tree is understood as containing the information on 1) the topology of the tree but not its edge lengths, and 2) for each edge, the identity of markers having undergone a mutation, in the sense of the previous paragraph. For any given phylogenetic tree t (in the previous sense), we can define $x_{ij}(t) = 0$ or 1 according to whether cell i carries a mutation at marker j (i.e., descends from an edge carrying a mutation at j) or not, based on tree t. The posterior of a tree t is a measure of agreement of the matrices Y and $X(t)$. More specifically, it is the probability of Y under the assumption that conditional on t, the variables $y_{ij}(t)$ are independent and the law of $y_{ij}(t)$ conditional on $x_{ij}(t) = \epsilon$ is Bernoulli with parameter p_ϵ, where $p_0 = r_{FP}$ is called a 'rate of false positive' and $1 - p_1 = r_{FN}$ is called a 'rate of false negative'. The results of the method are tested against synthetic datasets simulated under various assumptions, including conditions violating the perfect phylogeny assumption and compared to results obtained under other baseline methods. The method is extended to assign SNV to edges of the tree inferred by sitka. It is also applied to real datasets of single cell genomes of tumors.*

Main comments. *I concur with the comments of Reviewers 1 and 3, in particular:*

It would be good to improve the structure of the paper and expand some bits in order to make it readable by a wider audience (cf. comment of Reviewer 3). For example, the authors might like to expand the Introduction in order to have the reader better understand the context (low sequencing coverage but additional information coming from CNA), the specificities of the method, its main aspects and its applications (subclonal structure?), similarly as in my personal summary.

Response: Thank you for the suggestion. We have amended the introduction with three paragraphs, one providing background on the type of sequencing platforms motivating this work, one providing additional information on CNA, and one on applications.

Comment: *There are also several concepts and tools that should be defined more accurately (perfect phylogeny, overlapping/non-overlapping CNA, main principles of methods like UPGMA and the like, doublet/mouse cell/cycling cell, delta method, Sackin/Colless/Yule...).*

Response: We have improved the description of several concepts, including perfect phylogeny, overlapping/non-overlapping CNA, UPGMA, WPGMA, Neighbour Joining, HDBSCAN, MEDALT, MrBayes, medicc2, doublet cell, cycling cell, delta method, Sackin, Colless, Yule.

Comment: *There are also Supplemental figures that could be included in the*

main text (e.g., Supplemental Figures 1 to 4).
Response: We have moved the following supplementary figures to the main text: Supplementary Figure 1: moved to Main Figure 1; Supplementary Figure 2: moved to Main Figure 5; Supplementary Figure 3: moved to Main Figure 6; Supplementary Figure 4: moved to Main Figure 2.

Comment: *Following up on the previous point, a wealth of existing methods are cited in Introduction but their relation to sitka, as well as the difference/similarity with the benchmarked methods, should be better explained (maybe proposing a rough classification).*
Response: We have expanded the introduction section to better explain existing methods.

Comment: *Some other methods should be benchmarked, as mentioned by Reviewer 1.*
Response: We have added comparisons to two recently developed methods, benchmarking them against sitka on three real datasets. We have updated Figure 2d and section 2.2 to incorporate the new results.

Comment: *The claim that sitka relaxes "the independence assumptions required by existing phylogenetic methods" is not sufficiently well explained. Indeed, the method does not assume that copy number states evolve independently at different sites, but it assumes that the endpoints of CNA events occur independently, which may approximately hold for say the left endpoint, but not when combining both endpoints (see comment of Reviewer 3). Worse than that, it assumes independence of false positive/negative processes between lineages of different cells. It would be good to emphasize these aspects, to discuss the advantages and shortcomings of these assumptions (in Discussion) and also, as asked by Reviewer 3, to test violation of within-site independence, for example by assuming in synthetic experiments that sizes of CNA events are not exponentially distributed but e.g., always have the same fixed value or follow a heavy-tailed distribution.*
Response: Reviewer 3 and yourself are correct that our method ignores certain pairwise dependencies, and we agree this is a critical point to discuss. To address this point we have performed an additional set of experiments and highlighted this point in the discussion section. Please refer to our response to Reviewer 3 for a detailed account of the additional experiment.

Comment: *Could you maybe quantify the trade-off (mentioned line 58) between scalability/computational time and estimation accuracy ?*
Response: Exact quantification of the trade-off between computation time and estimation accuracy is difficult, for two reasons. First, ground-truth data in phylogenetic inference is scarce. Second, the models at the realistic end of this "scalability vs. accuracy continuum" are currently too costly to infer within a Bayesian framework, at least based on our parallel computing-based probabilistic programming tools. Perhaps in the future, distributed/MPI-based

Bayesian inference would allow us to investigate this, but we are not aware of user-friendly distributed computing probabilistic programming languages that would allow that at the moment. See also the related discussion later in this letter, under reviewer 1, point 5, second paragraph, where a similar trade-off is observed between MrBayes and the UPGMA method.

Comment: *Following up on the previous point (again), don't you think it would be more natural to model violation of 'perfect phylogeny assumption' by modeling directly the biological CNA process (gains and losses) as you do it verbally at the bottom of page 3 and in Supp Fig 3? In particular, I don't understand the IS violation procedure applied to the processed data (merging two columns): how do you do the merging and why does it mimic homoplasy? More generally, can you argue why you apply a lossy transformation to the data before analyzing it? At first sight, it looks like you lose a lot of information by replacing CNS by a binary variable telling whether contiguous bins have different CNS or not. In addition, the method assumes that the ancestral state of this binary variable is 0, which of course does not always hold in reality. Why can't you encode the data by the difference between CNS at two contiguous bins, so you don't lose the pseudo-independence of marker evolution at different marker sites but can keep the information on CNS and compute tree likelihood under the model of CNA evolution used in the simulations? Please discuss this.*

Response: We acknowledge that we lose information with application of the sitka transformation. This transformation is necessary for the computational feasibility of the likelihood, especially as required in Equations (3) and (4). In absence of this relaxation, the computational complexity of each iteration of the MCMC algorithm may no longer be bounded by $O(|C| + |L|)$. Indeed, the approach to efficiently compute the likelihood depends on binary latent variables with specific perfect phylogeny assumptions, and it is not clear how to generalize this calculation to models that keep track of the evolution of difference between CNS at two contiguous bins.

Comment: **Minor comments.**

General comment – isn't there sometimes a double meaning of the word 'locus'? (used both to denote a bin and a marker)

Response: We have updated the manuscript and changed the use of locus to marker or bin whenever the context is ambiguous.

Comment: *line 49 – The fact that likelihood-based methods perform statistically better than e.g., distance-based ones, should be supported by examples or references.*

Response: We have added three more references in our discussion of distance-based methods.

Comment: *line 104 – Please add that this procedure consists in passing from a type I to a type II tree. Anyway, I'm not sure you really need to explain the reader the difference between type I and type II trees.*

Response: We have updated the text to clarify this.

Comment: *line 130 – What do you mean by "sitka's performance degrades gracefully in the face of some of the key types of expected violation of the perfect phylogeny assumption"?*
Response: We agree that 'gracefully' was vague. We have changed this sentence to "Synthetic experiments show that sitka's performance decreases roughly linearly as a function of the rate of the key types of expected violation of the perfect phylogeny assumption (Fig 7-a,b)."

Comment: *line 222 – The Discussion section seriously needs to be fleshed out.*
Response: We have expanded the discussion section. Key additions include: the trade-offs of ignoring the pairwise dependencies between CNA end-points, performance of global vs the local noise models, trade-offs of the sitka-transformation.

Comment: *line 322 – Remove "disjoint"*
Response: Fixed.

Comment: *line 339 – Did you test the robustness of the method related to the upper bound of the support of the prior of false positive/negative rates?*
Response: We checked that the posterior of the FPR and FNR random variables were substantially away from the specified bounds for all real datasets in our study, namely OVA, SA501, and SA535. For example, below we show boxplots to summarize the posterior distribution for these quantities for the OVA dataset. Note that the FPR parameter posterior distribution concentrates tightly around 0.021, well away from the specified bounds 0.1. Similarly, the FNR parameter posterior distribution concentrates tightly around 0.098, well away from the bound of 0.5).

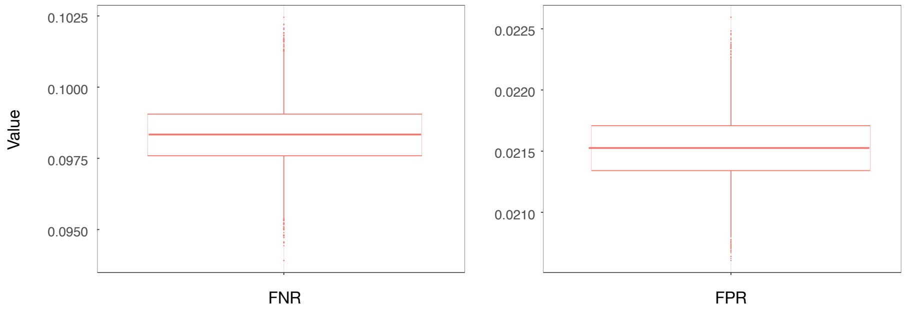

Figure 1

Comment: *line 359 – Please give a reference for the " 'rich gets richer' behaviour built-in into the prior, which is viewed as useful in many Bayesian non-parametric models" (see also comment by Reviewer 1)?*

Response: We added a relevant citation [1].

Comment: *line 391 – I like your definition of a Gibbs sampling algorithm ("an MCMC move with no rejection step"), but I am not sure it is very academic.*
Response: We removed that comment and added a reference to the following paper: Geman, Stuart, and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." IEEE Transactions on pattern analysis and machine intelligence 6 (1984): 721-741.

Comment: *line 429, eq (6) – Shouldn't you have $\pi(t^0, d\theta)$ rather than $\pi(t, d\theta)$? Maybe give a name to this argmin for future reference (see comment about lines 472–474).*
Response: Good catch, we have fixed this typo. We also updated the Eq (6) to denote the argmin with the notation $\tau_{\text{Consensus}}$.

Comment: *line 452 – Here and at some other places, you normalize scores by the score of the worst performing method. It seems weird because in the presence of a very poor-performing method for a given dataset, this will tend to overrate all alternative methods.*
Response: Our goal here is simply to make the results more comparable across the replication over synthetic datasets. We forgot to highlight the fact that we had included "Random" as one of the baselines, i.e., sampling a tree uniformly at random. Naturally, this is the worst performing method. We hope that this perspective of normalizing by a random reconstruction feels more intuitive. We have clarified this in the text.

Comment: *line 470 – Isn't the "best possible tree" just the true tree?*
Response: The best possible tree is derived from the true tree, but in general can be different. To understand why, recall that the tree generation process will simulate on each edge of the true tree a Poisson-distributed number of evolutionary events (Section 9.5.3). As a result, some edges can have zero associated evolutionary events. This means that even if we turned off all observation noise it would not be possible to recover some of these zero-event edges, i.e., they are in a sense unidentifiable. The process of producing the "best possible tree" essentially consists in collapsing these zero-event edges, hence forming a multifurcating reference tree. We have updated the text to provide additional explanations on the difference between the best possible and true tree.

Comment: *lines 472–74 – What is the difference between the "greedy estimator (GE) of Section 9.4.5" and the "trace search estimator (TSE) defined as a tree in the sampler trace that minimizes the sample L1 distance (Section 9.4.5)" ? After the latter definition, it seems to me that the TSE is given by Eq (6). Please give a mathematical formula for the estimator which is not defined by Eq (6) and specify which is which.*
Response: To see why TSE and GE are different, note that the former will always output one of the trees visited by MCMC while the latter can produce

a tree that has not been visited. Formally,

$$\tau_{\text{TSE}} = \operatorname*{arg\,min}_{t \in \{t^i\}} \sum_{t' \in \{t^i\}} L(t, t')$$

where $\{t^i\}$ denotes the set of trees that were sampled during the MCMC procedure. We have updated the text in Section 9.5.2 to clarify this.

Comment: *lines 478–484 and lines 496–498 – Please specify what is measured (RF distances? Normalized? Confidence intervals?).*
Response: The reported quantities are normalized RF distances along with standard errors. We have added a description in the manuscript.

Comment: *line 478 – Can you explain in Discussion why the global model can outperform the local model?*
Response: We made the observation of the global model performing better only in the context of TSE. In the context of GE, the global and local parameterizations performed similarly. Our recommendation to use the global parameterization stems from the facts that (1) both GE variants outperformed both TSE variants, and (2) that the global parameterization is computationally cheaper. We have updated the discussion to make this point.

Comment: *lines 509 – Add "of size s" (to "An island...")*
Response: Fixed.

Comment: *line 512 – Is there a reason why the violation rate thus estimated has anything in common with the violation rate defined in the synthetic experiments?*
Response: We acknowledge that the method used in that paragraph is heuristic and we have modified the text to emphasize this.

The intuition behind this heuristic is that a loss leaves a distinctive signature in the matrix $z = x - y$, where y is the data matrix, $x = x(t)$, and t is the consensus reconstruction. Consider for example the following figure:

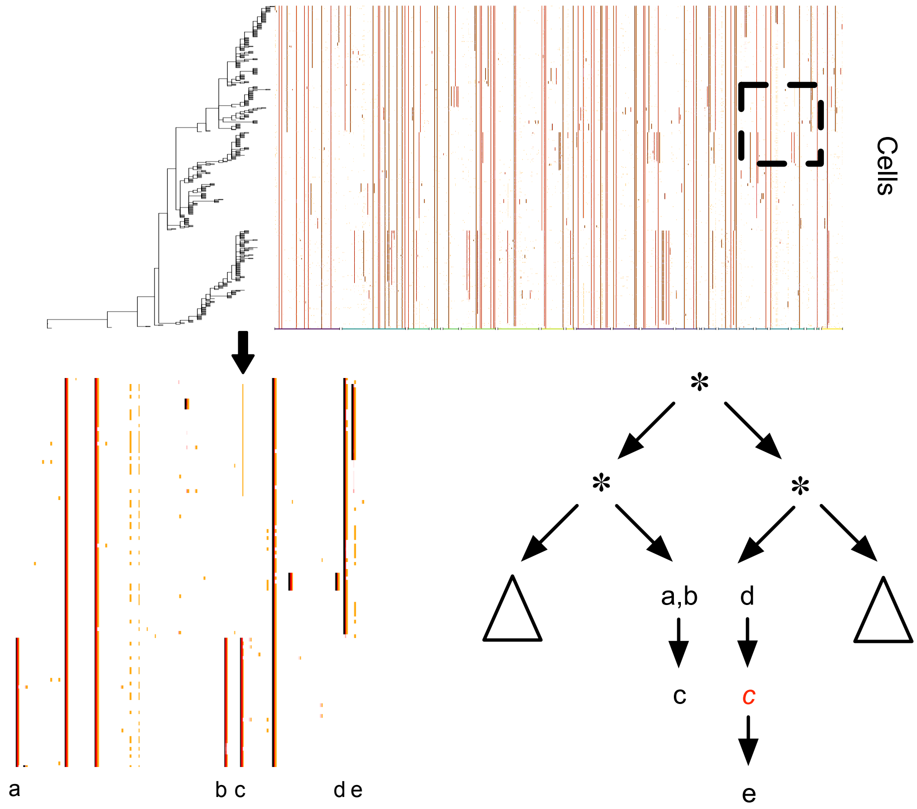

Figure 2

The bottom left shows an inset from the matrix in the top. Here, orange indicates whether the marker was observed in the input data, while black indicates whether the marker is present in the tree. The bottom right panel is a schematic of the tree topology for this part of the tree where the clade on the left is supported by markers a, b, and c, while the clade on the right is supported by marker d. However, in a violation of the perfect phylogeny assumption, marker c is also present in the clade supported by marker d. This signature manifests itself (in the bottom left panel, indicated by the black arrow) as a column of orange, unaccompanied by a black column. Our heuristics attempt to identify this signature automatically.

Comment: *line 552 - From my personal experience, the range of β for which Beta-splitting trees are interesting and realistic is (-2, 0) rather than (-1, 10).*
Response: The simulator [2] used in our experiments uses the generalized [3] Beta-Splitting model [4], which itself is different from the Beta-Splitting model of Aldous [5]. This distinction between models is mentioned in lines starting at 518 and 522.

The generalized Beta-Splitting model with beta range (-1, 0) is comparable to Aldous' Beta-Splitting model with beta range (-2, 0). Briefly, the generalized Beta-Splitting model is parameterized by two parameters α and β. In the case $\alpha = \beta$, we recover Blum-Francois's (BF) Beta Splitting-model. As such, the BF model itself is parameterized by a single beta parameter. The beta parameter in the BF model can take values > -1, whereas Aldous' Beta-Splitting model's β parameter can take values > -2. The authors [3] argue the expressivity of the BF model is quite general; the BF model covers a wide range of tree topologies. As β approaches -1, in the BF model, the realized trees become totally unbalanced; as β approaches infinity, realized trees become very balanced trees. In our experiments, simulated trees include balanced and imbalanced trees. Please see the **Supplementary Fig. 7** for examples of generated trees.

Comment: *line 566 - Why not follow the same procedure as previously? (Beta-splitting trees and, as in my last main comment, simulation of the biological CNA process)*
Response: That experiment is concerned with parameterizations of sitka and robustness to violations of assumptions. Thus we chose a simpler model to reduce the variability of results coming from other confounding factors that would arise from more complicated models.

Comment: *line 611–613 – Please give the mathematical formula defining $g_{.,j}$.*
Response: Updated in the text as follows. Let τ denote a rooted tree, u, one of its unlabelled internal nodes, and c one of its leaves. Let $\text{clade}(u)$ denote the clade corresponding to u, i.e., the set of leaves descendent from u. We define $g_{c,u}(\tau) = \mathbf{1}[c \in \text{clade}(\tau)]$.

Comment: *lines 622–624 – Please be more precise and elaborate notation to let the matrix o explicitly depend on its arguments (h, w, z?).*
Response: We have updated the text to make the definition of the matrix o more precise. In the process of doing that, we found that $o(..)$ is a redundant notation and replaced it by the essentially equivalent $h(..)$.

Comment: *line 669- – 'loci' should be 'locus' (twice).*
Response: Fixed.

Comment: *Supp Fig 2 – Please specify that the red nodes in (a) correspond from top to bottom to markers 2, 3, 1 in this order. Also if you feel it is important for the reader to understand the difference between type I and type II trees, it might be good to display a type I tree with more interior marker nodes on the same edge.*
Response: We have amended the caption to **Supplementary Fig.** 2 and added this clarification.

Comment: *Supp Fig 3 – "By the infinite site argument" is confusing (assumption vs approximation?)*

Response: We have updated the text to clarify this point as follows:

"If the infinite sites assumption holds, it is unlikely for the end-points of the two gain events to exactly match."

Comment: *Supp Fig 10 - Please specify that only Sackin and Colless indices are normalized and have positive values indicating more imbalance (it is the contrary for β). Do you have a sense why Sackin and Colless indices always give very similar values and why β is consistently estimated by $\approx$ -1?*
Response: Sackin and colless are similar metrics. The former is the sum of the depth of the leaves, while the latter is the sum of the absolute value of the difference between the number of leaves of the left and the right child of each internal node. We note that the normalized values of the two metrics are similar, but not identical.

We follow the implementation in [4] to examine the normalisation procedure:

$$\text{Sackin} := \text{INS}/\text{leaf}_{nb} - 2\sum_{j=2}^{\text{leaf}_{nb}} 1/j$$

$$\text{Colless} := \text{ICN}/\text{leaf}_{nb} - ((\sum_{j=2}^{\text{leaf}_{nb}} 1/j) - \ln(2))$$

where leaf_{nb} is the number of leaves in the tree and INS and ICN denote the unnormalised Sackin and Colless metrics respectively.

For the beta statistic β [5], the maximum likelihood estimates for SA501, OVA, and SA535 are -1.18, -1.31, and -1.33 respectively. We have normalized these values by dividing by the absolute value of their maximum (1.18) for ease of plotting. That is why the plotted numbers appear very close to -1.

2 Reviewer 1

We very much appreciate your helpful suggestions and comments regarding our paper. Below are our point-by-point responses to your suggestions and comments. The comments and questions are all included for convenience.

Comment: *In this manuscript, the authors introduce a novel, scalable method to infer phylogenies from single-cell whole-genome sequencing data based on copy number information. The algorithm is applied to three independent datasets and the goodness-of-fit compared to other methods. Possible violations of the model assumptions are discussed and put in context of real-world data. After tree inference, SNV data can be incorporated into the model prediction as well. The manuscript is very well written and the method appears to be fast and to perform favorably compared to other approaches.*

I would first like to commend the authors on the clarity of the majority of their manuscript and the high degree of detail. It was a pleasure to read it. I am further providing my detailed feedback and questions below.
Response: Thank you!

Comment: *1. Based on Eq. 2, it appears like the sampling probability of a vertex v is both proportional to the likelihood of each sub-tree, expressed by p(y—x(t),theta), as well as the number of possible sub-trees. The latter implies that vertices with many children are more likely to be sampled than trees with fewer children. Is this correct? Is this desired?*
Response: It is correct that the index set of the summand in (1) will grow with the arity of a node (number of children). This cannot be avoided as it follows from the structure of the space of multifurcating trees (at this step of the derivation, we are asking the question "what is the mass of possible trees that can be obtained with one edge insertion below v?"). In our context it is important to allow multifurcation since we are in a regime where there may not be enough markers to fully resolve all binary splits. Intuitively, it seems reasonable that the prior hence implicitly "encourages" resolving high-arity multi-furcations. If this behavior is not desired one could theoretically use a non-uniform prior over trees, however, depending on the details of the non-uniform prior this may complicate the design of marginalization for efficient MCMC sampling.

Comment: *2. How does the equation following l. 398, which posits that the probability $p(y_{c,l}|x_{c,l}, theta)$ of a vertex can be expressed as the product of probabilities of all its children, relate to the original definition of $p(y_{c,l}|x_{c,l}, theta)$ given after l. 322, according to which children and parent nodes are independent?*
Response: The equation after l. 398, p_v^b, denotes the initialization of a recursion relation that only holds at the leaves. At this point of the argument we have not explained yet the relationship with $p(y_{c,l}|x_{c,l}, \theta)$. The relationship is a bit involved and explained in the equations following "Putting it together.." So the initialization of this recursion does not contradict the independence statements in equations after l. 322.

Comment: *3. The transformation given in the equation after l. 413, which results in a product over k factors, contains all possibilities for edge insertions in sub-tree v, including the one in which no edge is inserted. Hence, vertices whose existing configuration already has a high likelihood are, counterintuitively, selected for edge insertion proportionately to this likelihood. I can imagine that, at best, this would slow the convergence of the algorithm, but there might be more deleterious consequences.*
Response: In all scenarios, an edge will be inserted, however when $\mathbf{b}$ is a vector with all elements equal to zero, then no cell will be selected to be moved under the newly inserted edge.

Comment: *4. Regarding the inference of the consensus tree (section 9.4.5),*

I am not sure I understand well Eq. 6. It appears the authors are using a generalized Bayes estimator by minimizing the posterior expected loss, as the loss function is weighted with the posterior distribution that is given after l. 365. Is this correct?
Response: Modulo the typo raised in the next point (thank you, it is now fixed), since the prior is proper, we are approximating the standard Bayes estimator (the terminology 'generalized Bayes estimator' is typically reserved for situations where the prior is improper). We have added one more step, starting the derivation from the definition of Bayes estimator and also simplified the notation by changing $\pi(t, d\theta)$ to $\pi(t, \theta)d\theta$.

Comment: *Second, it appears that then the parameter t in the argument of pi should be t'.*
Response: This is an astute observation by the reviewer and we have fixed this typo in Equations (6) and (7).

Comment: *Third, why did the authors choose to use the Bayes risk to determine the tree, especially since it appears that the priors for t and theta are anyway largely non-informative? Could they not just maximize the likelihood p(y—x(t), theta)?*
Response: We have added an experiment comparing our Bayes estimator to a Maximum Likelihood Estimator (MLE). The new results show that our Bayes estimator achieves a lower error rate compared to MLE and are described at the end of Section 9.5.2.

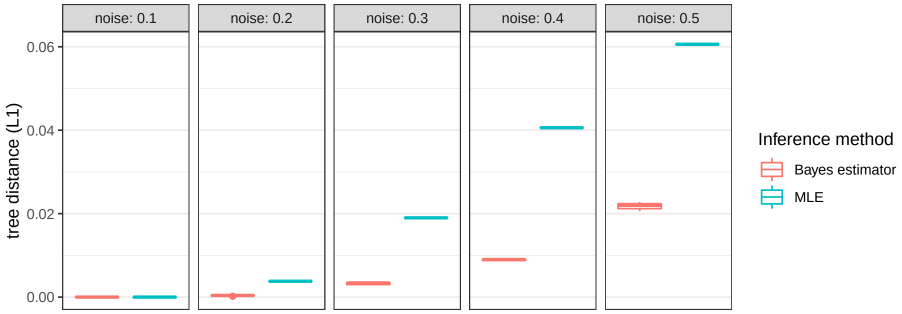

Figure 3

Comment: *5. In the benchmarking (Fig. 2d), I think the authors should compare their method also to other existing methods, in particular SCARLET from ref. [17] and MEDALT from ref. [18].*
Response: We have added comparisons to two recently developed methods, benchmarking them against sitka in three real datasets. For reasons described below, the two additional methods we have added are MEDALT and medicc2.

We have updated Figure 2d and section 2.2 to incorporate the new results. Note that on one of the datasets (SA501, the dataset with the largest number of cells), when provided with the integer copy number matrix as input, medicc2 did not finish running after 5 days. Similarly, on the same dataset MEDALT ran out of memory (we provided 144 GB of RAM memory). However, when we provided the sitka-transformation matrix as input, both methods finished running.

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.

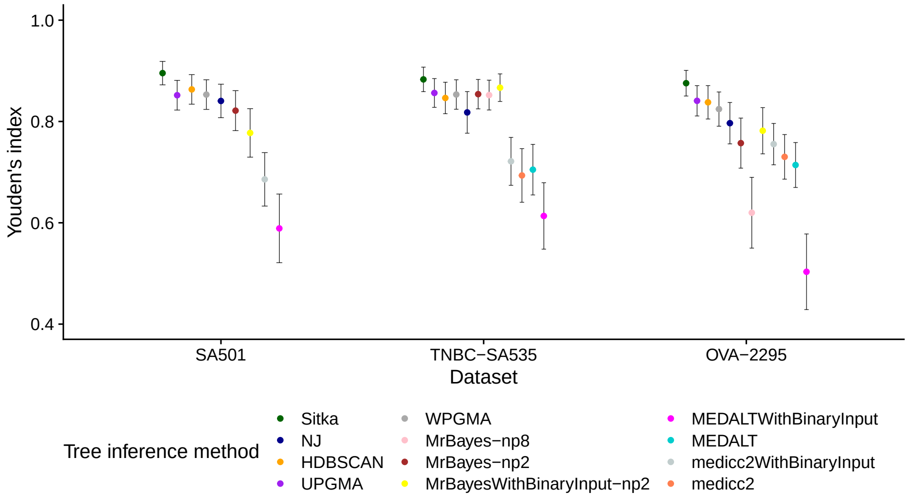

Figure 4

Comment: *In this regard, I also find it interesting that a simple hierarchical clustering method, UPGMA, shows such good performance, which is even largely better than that of MrBayes, when using Youden's index as a measure. Do the authors know why this would be so? Could it be that the performance measure is suboptimal?*

Response: We hypothesize that MrBayes does not converge in our available computational budget (several days). In line with this view, in the smallest dataset, SA535 that has only 493 cells, MrBayes with binary input is the second best performing method after sitka, and outperforms UPGMA. In the other two datasets that have more cells, MrBayes performs worse.

Comment: *6. Fig. 1f shows leaves without cells or markers. Why and how are these generated in the MCMC tree exploration scheme?*

Response: Fig. 1f shows zoomed-in insets from panel e. Each inset depicts a subtree, where the red diamonds are marker nodes and the blue circles denote single-cells. We have clarified this in the caption.

Comment: *7. Along similar lines, in l. 101 of the manuscript is stated that "Markers placed at the leaves are interpreted as outliers, for example measured CN change points that are false positives. We remove from the type I tree all marker nodes that are leaf nodes, i.e., markers that are not present in any cells." Would it not be possible to always place a marker that was observed in only one cell at the leaf corresponding to that cell?*
Response: Yes, this scenario is in line with the perfect phylogeny assumption and the corresponding marker node will be placed as a parent of the cell that harbors the locus. However, such an event is likely to be removed in the preprocessing.

Comment: *Also, why are such outliers still observed (e.g. CN change points that are false positives) if columns in y with relative density across cells less than 5% were removed (l. 314)?*
Response: It is true that in most scenarios (given more than 20 single-cells), an event that is only observed in one single-cell is removed in the preprocessing step. However, more prevalent outliers that pass the filtering preprocessing step may occur. For instance, take a region of the genome that due to sequencing artefacts presents multiple copy number events (all false positives). Such events can pass the preprocessing filter.

Comment: *8. In the computation of $p(y|x, theta)$ on p. 10, all entries (c,l) are assumed independent. However, change points should be correlated with some auto-correlation function with a decay rate proportional to the typical CN lengths, i.e. $Cov(y_{c,l}, y_{c,l'})! = 0$. Would it be feasible to incorporate this kind of information in the algorithm?*
Response: While it is conceptually possible to design a model incorporating dependencies among change points, it will negatively affect the computational runtime. More specifically, it is not obvious to us how the $O(|L| + |C|)$ upper bound cost of each MCMC iteration would be possible. It can be seen in Section 9.4.3 (Equations for ρ_v, in the bottom of page 13) that were $y_{c,l}$ not independent, the Gibbs conditional probability of selecting a marker v will no longer efficiently factorize.

Comment: *9. It appears like the FN and FP rates could be optimized instead of set to default values (0.5 and 0.1, respectively). Are these defaults informed or are they arbitrary?*
Response: In all experiments, the FN and FP rates are sampled jointly (the full Bayesian equivalent of optimizing over these parameters). The values 0.5 and 0.1 are prior hyperparameters and the actual FN and FP rates are inferred in posterior inference.

Comment: *10. Regarding the number of possible trees derived after l. 352,* *how come the second factor in the second line of the equation is $(|L|+1)^{|C|}$? Should there not be $(L+1)!/(L+1-C)!$ possibilities to assign $|C|$ cells to $|L|+1$ vertices?*

Response: The expression $(|L|+1)^{|C|}$ is correct. Note that the cells can be assigned to any loci, that is, a tree with all cells assigned to one locus is supported. Moreover, the vertices and cells are all labeled and unique, therefore, the assignment procedure amounts to selecting for each unique single-cell c, one of the $(L+1)$ unique loci (L traits under study and the virtual node denoting the root), which takes the form $(L+1)\cdot(L+1)\cdot = (L+1)^C$. The factorial form of the formulae $(L+1)!/(L+1-C)!$ would undercount the total number of possible unique trees.

Comment: *11. In l. 356ff., it is stated "This simple prior has a useful property: if a collection of say two splits are supported by m_1 and m_2 traits, then the prior probability for an additional trait to support the first versus second split is proportional to (m_1+1, m_2+1). Therefore, there is a "rich gets richer" behaviour built-in into the prior". How is this compatible with the prior being a uniform prior (cf. l. 353 and formula)?*

Response: The prior is uniform on the set of type I tree (Section 2.1), i.e., on the set of possible outputs of the two step process (Section 9.4.3). At the same time, it is possible to group trees into equivalence classes and look at the "induced prior" on these equivalence classes, i.e.,

$$\text{induced prior(class)} = \sum_{t \in \text{class}} \text{prior}(t).$$

Specifically when discussing the "rich gets richer property", we are considering the equivalence relation such that two type I trees t, t' are in the same equivalence class if and only if $f(t) = f(t')$, where $f(.)$ consists in transforming t into a type II tree and annotating each edge by the number of events on that edge. Since there are different numbers of type I trees in different equivalent classes, this means that the induced prior on these equivalence classes is non-uniform.

We apologize for the lack of details in the initial submission. We have added more details on that point in the manuscript.

3 Reviewer 2

Thank you very much for taking the time to read our manuscript.

Comment: *This paper develops a new method for phylogenetic modeling and Bayesian inference of cancer evolution that suggests being efficient when applied to tens of thousands of high-resolution genomes from single cell whole genome sequencing (scWGS).*

I think that the method clearly shows utility, but that it is not entirely clear whether it outperforms alternative approaches. This, however, is mentioned in the manuscript.
Response: We have added comparisons to two recently developed methods, benchmarking them against sitka in three real datasets. We have updated Figure 2d and section 2.2 to incorporate the new results. Note that on one of the datasets (SA501, the dataset with the largest number of cells), when provided with the integer copy number matrix as input, medicc2 did not finish running after 5 days. Similarly, on the same dataset MEDALT ran out of memory (we provided 144 GB of RAM memory). However, when we provided the sitka-transformation matrix as input, both methods finished running.

Comment: *The study of synthetic experiments helps readers to navigate the method and evaluated its impact and future utility, especially in light of cell removal due to contamination. I am intersted in seeing future applications of this method.*
Response: Thank you! Motivated by this, we have expanded the discussion of potential applications in the introduction.

4 Reviewer 3

We very much appreciate your helpful suggestions and comments regarding our paper. Below are our point-by-point responses to your suggestions and comments. The comments and questions are all included for convenience.

Comment: *The authors of the manuscript present a new method for reconstructing single cell phylogenies from previously inferred CNV data. Specifically, they propose a data transformation for CNV counts into discretized coarse grained markers of changes, based on which the phylogenetic reconstruction is performed efficiently. Importantly, the authors also propose a single point mutation calling method that conditions on the CNV based phylogenies to better resolve signal to noise problem. The method is compared to other state of art methods on three single cell datasets.*

The methods and algorithms are comprehensively presented.
Response: Thank you!

Comment: *In general, the manuscript would benefit from introducing more explanatory comments and brief motivations for introducing steps of the analysis, especially in the Results section (e.g. why do we need type I and type II trees, what is the benefit of using change points, or even definition of perfect phylogeny) so that non-expert readers can follow.*
Response: We have expanded the introduction and the discussion sections. Moreover, throughout the manuscript and in the supplementary materials, we

have added definitions to multiple concepts.

Comment: *However, I do have a major concern about the property of the sitka transformation and the effect it has on the phylogeny recontruction that the authors should adress. Each copy number variations comes with 2 breakpoints. The sitka transformation, to my understanding, ends up treating copy number changes along the chromosome as independent events, and, effectively, the markers of the beginning and the end of a copy number variation are not paired. What is the impact of this on the phylogenies? Are the pairs of breakpoints separated on the reconstructed phylogenies? If so, how distant they are? The authors should discuss this point and present the relevant statistics on empirical data.*

Response: The reviewer is correct that our method ignores certain pairwise dependencies, and we agree this is a critical point to discuss. To address this point we have performed an additional set of experiments and highlighted this point in the discussion section.

We now make this point the first one covered in our updated discussion. In particular, we emphasize the fact that this artificial "duplication" of the events having two input end-points can lead to the method being overconfident, i.e., outputting credible intervals that are smaller than they should be. This is partly a reason for focusing more on point estimates (consensus trees) in the present manuscript, which we expect are less affected by this phenomenon.

To further investigate this issue, we first make the observation that if we subset the sitka markers to keep only those where the copy number is increasing from left to right, we retain only one end point of each paired event. This creates a smaller set of independent markers $L' \subset L$. Next, we computed one sitka tree t based on all L loci (which includes ignored pairwise dependencies), and one sitka tree t' based on L' (a smaller set of independent loci). We then looked at the proportion of identical entries in the matrices $x(t')$ compared to $x(t)$, the latter subsetted to the columns in L'.

We performed the experiment described above on the S90 datasets (described in Section 9.5.3), and three noise regimes with increasing amounts of noise injected: (I) where step (ii) in Section 9.5.3 is skipped; (II) uniform noise parameters FPR and FNR drawn from uniform distributions on the intervals $(0.0005, 0.005)$, $(0.005, 0.015)$ respectively, doubling noise parameters drawn from a uniform distribution on $(0.015, 0.035)$ distribution, jitter noise parameters drawn from a uniform distribution on $(0.15, 0.35)$; (III) uniform noise parameters FPR and FNR drawn from uniform distributions on the intervals $(0.001, 0.01)$, $(0.01, 0.03)$ respectively, doubling noise parameters drawn from a uniform distribution on $(0.03, 0.07)$ distribution, jitter noise parameters drawn from a uniform distribution on $(0.3, 0.7)$. All results are averaged over 15 datasets.

In all three regimes we observed a large overlap between t and t', but this overlap is negatively correlated with noise: in regime (I) we observed a mean overlap of 0.99 (sd 0.004); in regime (II), a mean overlap of 0.97 (sd 0.009); in regime (III), a mean overlap of 0.76 (sd 0.18). The results support that in a low to moderate noise regime, it is reasonable to ignore violation of pairwise dependencies for the purpose of point estimation (consensus tree construction). In the higher noise regime (and/or for construction of credible intervals), it may be advantageous to build the two trees t and t'. We expect neither to systematically outperform the other, the trade-off being that t is built from more data but with independence violations, whereas t' is built from less data but without independence assumption violations. Our goodness-of-fit tests can be used to select one of these two trees for final output.

References

[1] Y. W. Teh et al. "Dirichlet Process." In: *Encyclopedia of machine learning* 1063 (2010), pp. 280–287.

[2] X. F Mallory et al. "Methods for copy number aberration detection from single-cell DNA-sequencing data". In: *Genome biology* 21.1 (2020), pp. 1–22.

[3] R. Sainudiin and A. Véber. "A Beta-splitting model for evolutionary trees". In: *Royal Society open science* 3.5 (2016), p. 160016.

[4] Nicolas Bortolussi et al. "apTreeshape: statistical analysis of phylogenetic tree shape". In: *Bioinformatics* 22.3 (2006), pp. 363–364.

[5] D. Aldous. "Probability distributions on cladograms". In: *Random discrete structures*. Springer, 1996, pp. 1–18.