molecular systems biology

Tree-ensemble analysis assesses presence of multifurcations in single cell data

Will Macnair, Laura De Vargas Roditi, Stefan Ganscha and Manfred Claassen

Review timeline:

Submission date:	12th July 2018
Editorial Decision:	24th September 2018
Revision received:	16th January 2019
Editorial Decision:	22nd February 2019
Revision received:	26th February 2019
Accepted:	26th February 2019

Editor: Maria Polychronidou

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 24th September 2018

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from two of the three referees who agreed to evaluate your study. Unfortunately, after a series of reminders we did not manage to obtain a report from reviewer #1. In the interest of time, and since the recommendations of reviewers #2 and #3 are quite similar, we have decided to proceed with these two reports. As you will see below, the reviewers acknowledge that the presented approach seems potentially useful for the field. They raise however a series of concerns, which we would ask you to address in a major revision.

I think that the recommendations of the reviewers are rather clear so there is no need to repeat the points listed below. All issues raised by the reviewers need to be satisfactorily addressed. As you may already know, our editorial policy allows in principle a single round of major revision so it is essential to provide responses to the reviewers' comments that are as complete as possible. Please feel free to contact me in case you would like to discuss in further detail any of the issues raised by the reviewers.

REFEREE REPORTS

Reviewer #2:

The authors present an algorithm for finding developmental branching points in high dimensional single cell data such as mass cytometry or scRNAseq. The main shortcoming in previous approaches that they see their algorithm as addressing is the lack of a method of assessing the validity of branch points. They also are concerned with the necessity of choosing a root or start point in some

algorithms
(Wishbone) and the "strong topological assumptions" of other algorithms such as Monocle 2.

The basic steps of the algorithm:
1. Data is transformed and dimensionally reduced with diffusion maps.
2. Density dependent down sampling is carried out as in SPADE.
3. Reference nodes (cells) are chosen so as to be evenly distributed in the cloud of cells.
4. The cells are partitioned by assigning each to the closest reference node.
5. Generate ensemble of trees. To generate one tree, randomly pick one cell from each partition and construct the minimum spanning tree (MST) using distances between chosen cells.
6. Data for the scoring of potential branch points are then gathered cutting each tree in the ensemble at each tree node and recording how many times each pair of cells ends up in the same branch for each node. This data is summarized in a matrix Bxij , with x being a node and I and j a pair of cells.
7. A score is then calculated for each node based on the matrix Bxij . The meaning of the scores is then assessed by comparison to scores for nodes in synthetic datasets without branch points. On this basis it is decided if the highest scoring node is a valid branch point.
8. Given that a branch point is found, a search is made for other branch points by applying TreeTop recursively to the branches emanating from the found branch point.

The innovative portions of this algorithm are steps 5 through 8 where the main contribution is the automatic scoring of potential branch points to avoid false positives. TreeTop seems to perform well on hierarchically branched synthetic data. Further to this, the authors tested Treetop on several previously published data sets: T cell maturation in the thymus from the Wishbone paper (reference 15), B cell maturation data from the Wanderlust paper (reference 9) and healthy bone marrow data from the viSNE paper (reference 22). While some of the results were mixed for these datasets based on expected biology (see major comments) for the T cell maturation data TreeTop recovers the expected branching to CD8 and CD4 as found by Wishbone.

Overall, TreeTop is a potentially compelling new algorithm that can quantitatively identify multiple branch points in high dimensional single cell datasets, however more evidence is needed to show that TreeTop is able to accurately and sensitively assess their validity. The manuscript would be significantly improved if its performance in this respect could be quantitatively compared to already existing data visualizations that have similar capabilities.

Major comments

- TreeTop finds no branch point in the B cell maturation data. This is problematic as a 3D PCA (first three PCA components) plot of this data with cells colored either for Kappa or Lambda clearly shows obvious well-separated Kappa and Lambda branches. Moreover, the bone marrow data in figure 2 seems underbranched compared to previous representations of that dataset and the expected biology. Together, these suggest at a minimum that the TreeTop algorithm is overly conservative in assessing branch points and that false-negatives could be an issue.

- Presumably, the generation of an ensemble of trees is what the authors mean by avoiding strong topological assumptions as in Monocle, which just generates one tree. On the other hand, the authors at one point state that the "embeddings found by Monocle identify exclusively trees, regardless of the topology of the data". This is confusing because TreeTop also only identifies trees, although an ensemble of them. Please clarify and elaborate.

- The synthetic data sets here are constructed to match numbers of cells, dimensionality of the single cell data, and standard deviation of the noise in the data. The synthetic data is constructed not to have branch points and various underlying dimensionalities: 0,1,2. Given that branch point identification is a key contribution, it is not clear that the synthetic data provides an adequate reference to assess the meaning of the scores for the actual data set of interest.

Minor comments
- The TreeTop force directed data visualization does not seem very clear in comparison to, for example, The Gephi forced directed visualization used in X-Shift.

Reviewer #3:

In this paper, Macnair et al. propose a new method TreeTop to identify and quantify branching point of biological processes from single-cell mass cytometry and RNA sequencing data. In addition, TreeTop also provides a graph-based method to visualize the learned ensemble of trees. TreeTop is able to overcome the limitations of existing approaches, including 1) the inability to detect the presence of a branch point; 2) the strong topological assumption 3) the supervised root point selection. The authors have tested TreeTop on previously published datasets depicting different biological processes including T cell maturation, B cell differentiation and hematopoiesis, as well as on synthetic datasets of different topologies proving their method's superiority and robustness. The authors also compared the performance of TreeTop to other two methods including wishbone and monocle for assessing the presence of branch points and global structure. In general, it is potentially useful in avoiding false positive branch points for many single-cell trajectory inference studies.
Major comments:
1. The methodology of branch identification in TreeTop, which mainly consists of density-based downsampling, building MST, constructing consistency matrix and deciding the final clusters, shares some similarities with the method Éclair (Giecold, G., Marco, E., Garcia, S.P., Trippa, L. & Yuan, G.C. Robust lineage reconstruction from high-dimensional single-cell data. Nucleic Acids Res (2016).) It would be worthwhile to compare it to Éclair and prove TreeTop's advantages over Éclair.
2. The authors agreed on the popularity of single-cell RNA-sequencing in studying single-cell transcriptional profiles. But in this paper, TreeTop is only tested on one RNA-seq dataset. Given the prevalence of scRNA-seq, I would recommend that the authors add more scRNA-seq analyses to make the experimental results more convincing.
3. In the preprocessing step, for single cell mass cytometry data, top diffusion components are used for T cell thymic maturation but not for the others. It would be helpful if the authors can explain how they decided whether to use diffusion map in the preprocessing steps.
4. Page 11, Paragraph 1. Last sentence 'The node with the largest branching score is the identified branch point.' To my understanding, each 'node' should contain many different 'points(cells)' within the corresponding Voronoi partition. Then which 'point/cell' should be used as the branch point? Or does the 'node' here have different meaning from the 'reference node'?
5. TreeTop needs to use a set of precomputed reference score distribution, which are dataset-specific and based on triangular synthetic data, to process new input data. This sounds more empirical and lacks solid proof. Given some more complex non-linear and concave non-branching structure (e.g. swiss roll), will it still work?
6. For the multi-layer branch point identification, instead of recursive application, is it possible to only run TreeTop once to get the multi-layer hierarchy based on the same branching score threshold? If not, what's the advantage of recursive application? Is it true that for the same point its branching score tends to get higher with the recursive division of initial tree? If that's the case, is it still fair to compare them directly after reassembling the subbranches since they are calculated in different configurations?
7. For TreeTop package, I ran it without success in matlab 2014b and got the following errors after strictly following the authors' tutorial. Hope the authors could solve this issue in their potential new version.

```
>> version
ans =

8.4.0.150421 (R2014b)
>> treetop_pre_run(input_struct, options_struct)
running pre-run analysis for TreeTop
1/6 Getting data
opening 1 files:
.
combining into one matrix
2/6 Plotting marginals of used markersUndefined function 'plot_fig' for input arguments of type
'matlab.ui.Figure'.

Error in treetop_pre_run>plot_marginals (line 41)
plot_fig(fig, plot_stem, options_struct.file_ext, fig_size);
```


Error in treetop_pre_run (line 16)
plot_marginals(all_struct, input_struct, options_struct)

Minor comments:
(1) The notation k is inconsistent in this paper. Page 10, k is the number of nodes. Page 14, k is the tree number. This can cause many confusions.
(2) The authors should mention the reason why monocle is only applied to the sample of 2000 cells instead of the full dataset.
(3) Page9, the last sentence in last paragraph 'This is problem is also largely resolved for the larger single cell RNAseq datasets measured with current droplet-based technologies.'. It's grammatically wrong.

1st Revision - authors' response 16th January 2019

Response summary to the referees' comments for manuscript MSB-18-8552

This is the response to the reviewer comments for the paper entitled "Tree-ensemble analysis assesses presence of multifurcations in single cell data".
We thank the reviewers for their overall positive evaluation of the novelty and significance of our contribution, as well as for their constructive suggestions. We have identified and addressed the following main issues raised by the reviewers:

1. **The considered synthetic reference data may not be an adequate reference for assessment of bifurcation presence in real data, and may result in overly conservative branch point identification.**

   Branching processes in single-cell data are complex patterns. Many statistical significance tests use permutations of the experimental data, however we found that assessing the presence of branch points with permutations results in over-reporting of branch points. We therefore designed synthetic reference data for branch point assessment which includes non-branching structures. We demonstrate that branch point assessment based on our approach leads to correct identification of branch point presence as well as absence for a range of examples of branching and non-branching processes.
   TreeTop was specifically designed to show some conservativity, as a benefit to users, in contrast to algorithms such as Wishbone (which always report branches). We sought to reduce spurious reports of branching which would lead to wasted time and resources for additional validation experiments. We have demonstrated that TreeTop is able to identify known branch points in multiple datasets, and therefore believe that TreeTop is appropriately, rather than overly, conservative.

2. **TreeTop should be demonstrated on more single cell RNA-seq datasets.**

   We agree with the reviewer that given the ever-increasing use of single cell RNA-seq technologies, a more comprehensive demonstration of TreeTop on this type of data would be beneficial. We have therefore revised the manuscript to include analysis of an additional dataset sampled from hematopoiesis (see new Figure 3 in revised manuscript).

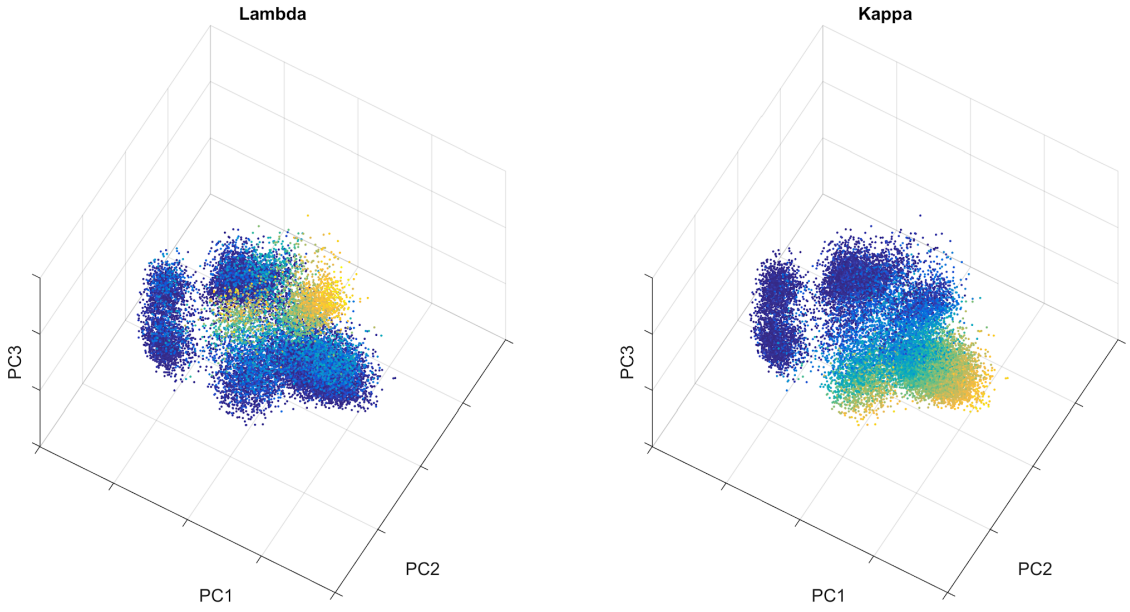
The individual comments are discussed in turn below.

Reviewer 2

2.1 "TreeTop finds no branch point in the B cell maturation data. This is problematic as a 3D PCA (first three PCA components) plot of this data with cells colored either for Kappa or Lambda clearly shows obvious well-separated Kappa and Lambda branches."

Reviewer 2 is concerned that our analysis of maturing B cells (Supp Fig 3, original manuscript) may be overly conservative. Maturing B cells are known to express either kappa or lambda light chains. Reviewer 2 argues that these separate fates are clearly distinguishable in the dataset in question, however we disagree with the reviewer on this conclusion. In addition, the results from TreeTop

applied to this dataset suggest weak evidence of branching, which is entirely consistent with biological expectations of this dataset.
In applying TreeTop to the B cell maturation data, we followed the analysis in the original publication, which evaluated dissimilarity of cells with cosine distance (rather than Euclidean or L1 distance). Following the reviewer's suggestion, we have plotted the first 3 PCA components for this data (see below). We observe clear *clusters* corresponding to high Kappa and high Lambda cells, but do not agree that these form clearly separate *branches*; without prior biological expectations, this is not a clear conclusion from the PCA plot.



However, we have updated the application of TreeTop in the manuscript to be based on the first 10 PCs (as these account for 90% of explained variance) of the B cell data using L1 distance (TreeTop's default). TreeTop identifies the kappa/lambda clusters noted by the reviewer, however the confidence score for branching in the dataset is just below the cutoff used by TreeTop (updated Appendix Figure S3, p27 in revised manuscript). We interpret this as weak evidence in favour of a branching process, a conclusion which might change if the data was sampled from the full B cell differentiation process (i.e. with greater sampling of the Kappa/Lambda-committed cells).
This example demonstrates the utility of one of the outputs from TreeTop, the branching score. The default score threshold would report the evidence for the bifurcation of the kappa/lambda clusters as just below the threshold for branching; the user could then identify this borderline situation and follow up if applicable.

2.2 "Moreover, the bone marrow data in figure 2 seems underbranched compared to previous representations of that dataset and the expected biology."

The bone marrow data in TreeTop Figure 2 is taken from the healthy control sample shown in Figure 6d of Amir *et al.* (Amir et al. 2013). The cell types identified in this data were: progenitor cells, T cells, CD20+ B cells, CD20- B cells, monocytes, NK cells and ungated cells. This dataset does not include markers which allow CD4+ and CD8+ T cells to be distinguished, which results in one fewer branch point. Our analysis does not clearly separate CD20+ and CD20- B cells, but these cells are also not clearly separated in the original paper. Our analysis is therefore consistent with previous representations of this dataset.

2.3 "Together, these suggest at a minimum that the TreeTop algorithm is overly conservative in assessing branch points and that false-negatives could be an issue."

TreeTop provides a score of confidence for identified branch points. This score is designed to be rather conservative than too optimistic, which assists users by reducing possible investigations of false positive branch detections. This feature is in contrast to currently available algorithms which always report branches, without any score of confidence, leaving the calling of bifurcations a subjective user decision. Additional experiments are costly, and therefore we sought to reduce spurious reports of branching which would lead to wasted time and resources. Even where TreeTop reports no evidence of branching, Wishbone reports multiple branches in all cases, and therefore

cannot be used for discovery of new branches (p3, para 2 and Appendix Figure S1 in revised manuscript). This demonstrates that Wishbone is insufficiently conservative. We also showed that TreeTop is able to identify known branch points in multiple datasets, demonstrating that TreeTop is not overly conservative in these cases. Our analysis above shows that there is only weak evidence of a branch point in the B cell data, as demonstrated by TreeTop. Taken together, these results show that TreeTop is appropriately, rather than overly, conservative.

2.4 "Presumably, the generation of an ensemble of trees is what the authors mean by avoiding strong topological assumptions as in Monocle, which just generates one tree. On the other hand, the authors at one point state that the "embeddings found by Monocle identify exclusively trees, regardless of the topology of the data". This is confusing because TreeTop also only identifies trees, although an ensemble of them. Please clarify and elaborate."

TreeTop produces both visualizations and a relative branching score, both based on the ensemble of trees.

For visualization, the ensemble of trees is summarized in a 'union' graph, which is effectively a superposition of all the trees in the ensemble (see Methods section *Ensemble of trees visualization*, p12 revised manuscript, for details on additional pruning of low-frequency edges). This union graph typically is not a tree and can contain cycles. To illustrate this point, consider a dataset sampled from the circumference of a circle (as in the third column of Appendix Figure S1). Here, Monocle fits a tree with no branches, but with a cutpoint at some point around the circle (Appendix Figure S1c). Each of the individual trees sampled by TreeTop also must include a cutpoint, at different points around the circumference, but taken together the topology they identify is correct (Appendix Figure S1d).

Scoring of nodes as potential branch points is based on analysis across the ensemble of trees (see section *Identification of branch points* in revised manuscript). Analysis across the ensemble, instead of a single tree alone, enables identification of consistent non-spurious branching structure. For a given node, each tree is cut at that node, separating that tree into branches. The branching score reflects large and consistent branches across the ensemble. In the case of the circle, the branches would not be consistent across the ensemble. Additionally, the branching score is based on the size of the third largest branch (as at least three branches are necessary to make a branch point). In this example, any third branches only result from noise in the data, resulting in low scores.

2.5 "The synthetic data sets here are constructed to match numbers of cells, dimensionality of the single cell data, and standard deviation of the noise in the data. The synthetic data is constructed not to have branch points and various underlying dimensionalities: 0,1,2. Given that branch point identification is a key contribution, it is not clear that the synthetic data provides an adequate reference to assess the meaning of the scores for the actual data set of interest."

Stated simply, the task which TreeTop addresses is to decide which of the following statements is true: this dataset contains a branching point, or this dataset does not contain a branching point. In principle, making this decision requires knowing the distribution of *all possible* non-branching datasets. Defining this distribution is a fundamentally difficult, poorly-defined problem, even if we make reasonable assumptions regarding measurement noise (meaning that we could exclude extremely contrived, biologically unrealistic examples constructed purely to result in high branching scores), continuity of data, and biological plausibility.

Standard statistical approaches address the problem of an unknown null distribution by permutations of the input dataset. However, permutations of the input data are not appropriate for assessing the presence of branch points, as permutations result in a complete loss of structure in the data: in addition to loss of structure induced by branch points, simpler structure such as that resulting from a dynamical process is lost. A consequence of this is that using permutation-based testing, structures in non-branching datasets would be reported as branch points. We therefore conclude that permuted input data is an incomplete approximation of the distribution of non-branching datasets, leading to a high rate of false positive branch point discoveries (see revised manuscript Appendix Figure S15, and section *Choice of synthetic reference data topologies for reference score distributions*, p15 paragraph 2).

We therefore sought to derive synthetic reference datasets which were non-branching, but resulted in the highest possible scores, meaning that when applied to real data we would minimize the number of false positive branch points reported. We considered simple, connected non-branching topologies, namely embeddings of simple, non-branching, low-dimensional manifolds in higher-dimensional space, with Gaussian noise. We showed that increasing the dimensionality or increasing the number of points considered in the synthetic distribution results in lower scores (see revised manuscript Appendix Figure S16, p46). This informed the use of the selected reference topologies,

which therefore exclude the widest range of non-branching datasets. To ensure that the topologies were appropriate to compare to a given input dataset, we calculated scores for the defined topologies for synthetic datasets whose data parameters (e.g. number of cells, dimensionality, number of reference cells) matched the input data.
In summary, defining the distribution of all non-branching, biologically plausible datasets is a difficult, unsolved problem. We have sought to address this gap by defining and identifying high-dimensional datasets with non-branching structure, which result in the highest possible branching scores, therefore minimizing possible false positive results. For accurate branch point identification in a new input dataset, we require that branch scores exceed all scores observed in non-branching datasets. We acknowledge that this procedure makes the assumption that we have considered all possible non-branching datasets for comparison. We have made an extensive empirical effort towards fulfilling this assumption and have demonstrated correct identification of presence as well as absence of branch points in synthetic and mass cytometry datasets. Assessment of branch point presence is a new, difficult and so far unaddressed problem to solve, and we present here a viable solution towards resolving it.

2.6 "The TreeTop force directed data visualization does not seem very clear in comparison to, for example, the Gephi force-directed visualization used in X-Shift."

X-Shift was developed specifically as a tool for visualization, and while TreeTop includes a visualization component, its primary focus is branch analysis. In principle, the Gephi force-directed visualization could be applied to the graph learned by TreeTop, or equally, the branching scores learned by TreeTop could be displayed over the k-nearest neighbours graph used by X-Shift. (The outputs from running TreeTop allow both of these possibilities in the following files: the scores for each node in *[RUN_LABEL] branching scores.txt*, the union graph learned in *[RUN_LABEL]_freq_union_tree.mat*, and the locations of the reference cells in *[RUN_LABEL]_mean_used_markers.txt*.)

Reviewer 3

3.1 "The methodology of branch identification in TreeTop, which mainly consists of density-based downsampling, building MST, constructing consistency matrix and deciding the final clusters, shares some similarities with the method Éclair (Giecold et al., Nucleic Acids Res (2016).) It would be worthwhile to compare it to Éclair and prove TreeTop's advantages over Éclair."

We downloaded Eclair and attempted to run it on some sample data. However, we were unable to get it to run successfully, and the lead author no longer works in academia. Researchers comparing trajectory analysis packages were also unable to successfully run Eclair ((Saelens et al. 2018), p2).

3.2 "The authors agreed on the popularity of single-cell RNA-sequencing in studying single-cell transcriptional profiles. But in this paper, TreeTop is only tested on one RNA-seq dataset. Given the prevalence of scRNA-seq, I would recommend that the authors add more scRNA-seq analyses to make the experimental results more convincing."

We have applied TreeTop to the Paul *et al.* dataset (see new Figure 3, panels a-c in revised manuscript). This comprises single cell RNA-seq data from 2730 developing myeloid cells, labelled as follows: granulocyte-macrophage progenitor (GMP), megakaryocyte-erythroid progenitor (MEP), erythrocytes (Ery), dendritic cells (DC), monocytes (Mo), basophils (Baso), neutrophils (Neu), eosinophils (Eos), megakaryocytes (Mk) and lymphocytes (Lymph). Here, TreeTop finds a branch point (comprising primarily of MEPs) which separates erythrocytes, megakaryocytes and other cells, then a further branch point separating these cell types. This is consistent with findings by other authors, for example (Perié et al. 2015).

3.3 "In the preprocessing step, for single cell mass cytometry data, top diffusion components are used for T cell thymic maturation but not for the others. It would be helpful if the authors can explain how they decided whether to use diffusion map in the preprocessing steps."

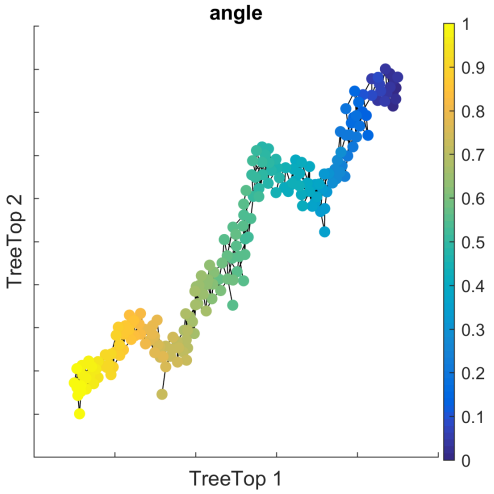
There are many possible methods for reducing the dimensionality of single cell data. Given the wide range of processes from which they are sampled, we do not believe that is sensible to specify a universal recipe for upstream analysis. TreeTop is compatible with any selected pre-processing. We would advise trying multiple dimensionality reduction techniques to identify the one which best reflects prior biological knowledge about the data, and potentially also running TreeTop using each set of pre-processing options. In this specific case, the cells corresponding to maturing CD4+ and CD8+ T cells were much more clearly separated in the diffusion map components, than in the PCA components.

3.4 Page 11, Paragraph 1. Last sentence 'The node with the largest branching score is the identified branch point.' To my understanding, each 'node' should contain many different 'points(cells)' within the corresponding Voronoi partition. Then which 'point/cell' should be used as the branch point? Or does the 'node' here have different meaning from the 'reference node'?

We thank the reviewer for identifying this lack of clarity. A node contains multiple cells within the Voronoi partition. Our phrasing here was insufficiently clear, and would have been better phrased as "... is the identified branch node"; this is amended in the revised manuscript. We believe identifying a group of cells as branch points makes the most biological sense (rather than a single cell). (p14, para 4 in revised manuscript)

3.5 "TreeTop needs to use a set of precomputed reference score distribution, which are dataset-specific and based on triangular synthetic data, to process new input data. This sounds more empirical and lacks solid proof. Given some more complex non-linear and concave non-branching structure (e.g. swiss roll), will it still work?"

TreeTop branch point analysis is based on triangular synthetic data, since this topology has shown to be the most confounding compared to other considered non-branching topologies.
Further, since TreeTop is based on a neighborhood graph structure (ensemble of trees), it is unaffected by characteristics of the dataset which do not affect the underlying topology (i.e. which do not change the neighbourhoods of cells). Non-linear, concave or other structures which do not have branching, still do not have branching topologies.
As suggested by the reviewer, we have applied TreeTop to a 10-dimensional swiss roll dataset as an empirical confirmation of this point. The plot below shows the results of applying TreeTop, annotated by the angle around the Swiss roll; here, TreeTop recapitulates the known topology and does not report any branching. This dataset is included in the example data included on the TreeTop GitHub page.



3.6 "For the multi-layer branch point identification, instead of recursive application, is it possible to only run TreeTop once to get the multi-layer hierarchy based on the same branching score threshold? If not, what's the advantage of recursive application? Is it true that for the same point its branching score tends to get higher with the recursive division of initial tree? If that's the case, is it still fair to compare them directly after reassembling the subbranches since they are calculated in different configurations?"

The branching score calculated by TreeTop is based on average sizes of any consistent branches at a given point. In a dataset with a hierarchy of branch points, branch points lower in the hierarchy will by definition have smaller branches associated with them. This means that one threshold cannot be used to detect all branches, although each branch point will be a local maximum of branching scores. The runs of TreeTops for different subbranches use comparison datasets with appropriate numbers of cells, making the scores comparable.

3.7 "For TreeTop package, I ran it without success in MATLAB 2014b and got the following errors after strictly following the authors' tutorial. Hope the authors could solve this issue in their potential new version."

We apologize for this, and thank the reviewer for supplying the error log. Our testing of the package was clearly not sufficient! This was an issue with required subfolders not being automatically on the path, and is now fixed.

3.8 "The notation k is inconsistent in this paper. Page 10, k is the number of nodes. Page 14, k is the tree number. This can cause many confusions."

Thank you for spotting this typo, now corrected. (p18 of revised manuscript, section *TreeTop Pseudocode*)

3.9 "The authors should mention the reason why monocle is only applied to the sample of 2000 cells instead of the full dataset."

Monocle becomes slow for larger datasets. We have included a note to this effect in the revised manuscript. (p19 para 5 in revised manuscript)

3.10 "Page 9, the last sentence in last paragraph 'This is problem is also largely resolved for the larger single cell RNAseq datasets measured with current droplet-based technologies.'. It's grammatically wrong."

We thank the reviewer for spotting this typo, which is now corrected. (p3, para 2 in revised manuscript)

2nd Editorial Decision 22nd February 2019

Thank you for sending us your revised manuscript. We have now heard back from the two referees who were asked to evaluate your study. As you will see below, the reviewers are satisfied with the modifications made and they think that the study is now suitable for publication.

Before we formally accept your study for publication we would ask you to address the following minor issues.

REFEREE REPORTS

Reviewer #2:

Based on the revision and response the authors give a good explanation of the issues involved in devising a methodology to determine the validity of potential branch points.

We acknowledge that the problem is very difficult and they have made a good worthwhile initial contribution to the solution of this problem. While the method may be somewhat conservative, since they do report scores of potential branch points, users may judge marginal cases for themselves. For instance, this is the case for the Kappa/Lambda branching in the B cell data, discussed in section 2.2, where they point out that their method does indicate the marginal possibility of a branch point. Their answers in 2.2 and 2.4 about branching and topologies are also reasonable.

Overall, the work in the paper presents a reasonable contribution to the analysis of branch points in single cell data and represents a unique contribution to the growing field of single cell trajectory analysis.

Reviewer #3:

In the revised manuscript, the authors have well addressed my concerns sufficiently. Great work. I would recommend it for publication.

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Corresponding Author Name: Manfred Claassen
Journal Submitted to: Molecular Systems Biology
Manuscript Number: MSB-18-8552

Reporting Checklist For Life Sciences Articles (Rev. June 2017)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

A- Figures

1. Data

The data shown in figures should satisfy the following conditions:

- the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
- figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
- graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
- if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified
- Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

2. Captions

Each figure caption should contain the following information, for each panel where they are relevant:

- a specification of the experimental system investigated (eg cell line, species name).
- the assay(s) and method(s) used to carry out the reported observations and measurements
- an explicit mention of the biological and chemical entity(ies) that are being measured.
- an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.

- the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
- a statement of how many times the experiment shown was independently replicated in the laboratory.
- definitions of statistical methods and measures:
  - common tests, such as t-test (please specify whether paired vs. unpaired), simple $\chi 2$ tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - exact statistical test results, e.g., P values = x but not P values < x;
  - definition of 'center values' as median or average;
  - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

In the pink boxes below, please ensure that the answers to the following questions are reported in the manuscript itself. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable). We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?	NA
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	NA
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	NA
3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.	NA
For animal studies, include a statement about randomization even if no randomization was used.	NA
4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.	NA
4.b. For animal studies, include a statement about blinding even if no blinding was done	NA
5. For every figure, are statistical tests justified as appropriate?	NA
Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.	NA
Is there an estimate of variation within each group of data?	NA
Is the variance similar between the groups that are being statistically compared?	NA

C- Reagents

USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repo

http://grants.nih.gov/grants/olaw/olaw.htm
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm
http://ClinicalTrials.gov
http://www.consort-statement.org
http://www.consort-statement.org/checklists/view/32-consort/66-title

http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tun

http://datadryad.org

http://figshare.com

http://www.ncbi.nlm.nih.gov/gap

http://www.ebi.ac.uk/ega

http://biomodels.net/

http://biomodels.net/miriam/
http://jjj.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right).	NA
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	NA

* for all hyperlinks, please see the table at the top right of the document

D- Animal Models

8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA
10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.	NA

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	NA
12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.	NA
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	NA
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	NA
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	NA
16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list.	NA
17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines.	NA

F- Data Accessibility

18: Provide a "Data Availability" section at the end of the Materials & Methods, listing the accession codes for data generated in this study and deposited in a public database (e.g. RNA-Seq data: Gene Expression Omnibus GSE39462, Proteomics data: PRIDE PXD000208 etc.) Please refer to our author guidelines for 'Data Deposition'. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Functional genomics data e. Proteomics and molecular interactions	NA
19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).	NA
20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right).	NA
21. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biomodels (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information.	Code for TreeTop is open source and is available at https://github.com/wmacnair/TreeTop.

G- Dual use research of concern

22. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could.	NA