Cell Genomics Transparent Peer Review Record

Unravelling dynamically-encoded latent transcriptomic patterns in pancreatic cancer cells by topic modelling

Yichen Zhang, Mohammadali (Sam) Khalilitousi, and Yongjin P Park

Summary

Initial submission:	Received : 3/11/2023
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 4/14/2023 Revision received : 5/27/2023
Second round of review:	Number of reviewers: 2 Accepted : 7/31/2023
Data freely available:	Yes
Code freely available:	Yes

Cell Genomics Transparent Peer Review Record

Referees' reports, first round of review

Reviewer #1: Comments enter in this field will be shared with the author; your identity will remain anonymous.The Authors present two Bayesian topic models (BALSAM and DeltaTopic), based on deep learning approaches, to discover short-term RNA velocity dynamics from spliced and unspliced single-cell RNA-seq reads.

Overall, the article is well structured and clearly presented, and the methods proposed are interesting, and well described.
The Authors validate their approaches on real data.

I have some concerns, mainly about:
- the distribution of the statistical software tool (major comment 1);
- the benchmark (major comments 2-4).

I suggest the article to undergo major revisions.

Simone Tiberi, The University of Bologna

Major comments:
1) The main aim of the study is to present a novel statistical software tool, distributed on python.
However, the GitHub repo is a collection of poorly documented scripts.
In my view, the software must be well documented and a clear example usage should be added (for a notable example, see scVelo's distribution: https://scvelo.readthedocs.io/en/stable/).

Also, the paper briefly mentions the availability of the tools, while to me this should be a key aspect to highlight (e.g., "BALSAM and DeltaTopic are freely distributed as … python scripts, at …link… .").

2) The methods are convincingly tested on real data.
Nonetheless, the benchmark should also include a simulation, to test method's performance against a ground truth.

3) The methods are not benchmarked against any other competitor.
A direct comparison with other methods is probably not possible because, I believe that, no other tool provides exactly the same output as BALSAM and DeltaTopic.
Nonetheless, it would be important to explain how the proposed methods differ from similar tools provide, and to show how results differ in real data.

4) A computational benchmark of runtime and RAM should be added.

Minor comments:
1) Line 7:
"[…] by taking the difference between the spliced and unspliced counts[…]"
I think the word "difference" is misleading, as it points towards a simple mathematical subtraction, which is not what RNA velocity tools consider.

2) Line 53: Q/C not defined.

3) Lines 84-85 mention dropout, however this used to be a concern in early single-cell RNA-seq technologies.
It is generally believed not to be an issue in droplet single-cell RNA-seq technologies.

4) Lines 86-89:
"For instance […] values (Fig. 1B on the right)."
The Authors present the 3 cases in Figure 1B as examples of "sparsity due to technologies".
However, I disagree with this view: spliced and unspliced abundances can differ; this does not imply a dropout event.

5) Related to the above comment, in the caption of Figure 1, the Authors argue that:
"The red dashed line indicates steady state, where the spliced and unspliced gene is in the same amount. "
To me, this is inaccurate: spliced and unspliced abundances can also differ significantly at steady state (e.g., depending on the degradation rates).

6) Methods: "Dieng and coworkers" I think should be followed by (year) and [citation].

7) Methods:
"(1) We introduces a Bayesian hierarchical prior […]"
Point (2) is on a separate line (bullet point style), so I think (1) also should.

8) Methods:
"with respect the variational parameters".
I think that "to" is missing after "with respect".

Reviewer #2: The authors use a deep variational topic model to learn the common and additional latent space for spliced and unspliced transcripotomic data. The model is mathematically sound and the findings are pretty interesting. I only have a few minor comments.

1. It would be nice to compare the BALSAM method to other types of latent space decomposition method like NMF. I know one is a probabilistic model and another is a matrix factorization. I am curious about what's the advantage of the probabilistic one. I think computational wise, NMF should be much faster.

2. I am curious about the broad application of the BALSAM method, so it would be helpful to try it on some generic transcriptomic data and see how the latent space it produced can compare to the others in terms of dimension reduction and clustering performance etc.

3. It would be helpful if the running time of the deltatopic method can be reported.

4. Some typos: line 71 gell-topic -> cell-topic; appendix: namely and xi -> namely and ξ.

Authors' response to the first round of review

We found all the points raised by the reviewers valid and strove to address them to our best.

Reviewer #1

The Authors present two Bayesian topic models (BALSAM and DeltaTopic), based on deep learning approaches, to discover short-term RNA velocity dynamics from spliced and unspliced single-cell RNAseq reads. Overall, the article is well structured and clearly presented, and the methods proposed are interesting, and well described. The Authors validate their approaches on real data. I have some concerns, mainly about: the distribution of the statistical software tool (major comment 1); the benchmark (major comments 2-4).

I suggest the article to undergo major revisions.

Simone Tiberi, The University of Bologna

We appreciate Dr. Tiberi for the careful and meticulous evaluations and thoughts, and efforts. First of all, we are very happy to know that our paper was read well enough to invoke interest among state-of-the-art scientists. We addressed the major and minor points in the following:

Major comments:

1) The main aim of the study is to present a novel statistical software tool, distributed on python. However, the GitHub repo is a collection of poorly documented scripts. In my view, the software must be well documented and a clear example usage should be added (for a notable example, see scVelo's distribution: https://scvelo.readthedocs.io/en/stable/).

We built our work on torch and scanpy since both are frequently used in single-cell genomics and machine learning analysis. Taking this point seriously, the first author (Yichen Zhang, a PhD student) set up the first version of documentation websites, https://deltatopic.readthedocs.io/en/latest/, which we hope to serve as a focal point in communication with researchers. Since we don't think of the current version as the final, complete product, we will keep on updating the software and documentation as needed.

Also, the paper briefly mentions the availability of the tools, while to me this should be a key aspect to highlight (e.g., "BALSAM and DeltaTopic are freely distributed as … python scripts, at …link… .").

The software library was now made installable via either pypi repository or our custom GitHub library.

2) The methods are convincingly tested on real data. Nonetheless, the benchmark should also include a simulation, to test method's performance against a ground truth.

We also felt the need fo additional benchmark analysis. So, we added a new sub-section at the end of Result, dedicated to performance comparison based on forward simulations with gold standard answers.

3) The methods are not benchmarked against any other competitor. A direct comparison with other methods is probably not possible because, I believe that, no other tool provides exactly the same output as BALSAM and DeltaTopic. Nonetheless, it would be important to explain how the proposed methods differ from similar tools provide, and to show how results differ in real data.

We agree with the reviewer's suggestions. We added a new section dedicated to benchmark analysis; there, we compared other relevant methods. We believe this new section has markedly improved this paper, leading to richer discussions on integrative data analysis.

4) A computational benchmark of runtime and RAM should be added.

We added a paragraph describing the computational resources that our training algorithms consumed. We acknowledge our approach is not ideal in this sense as most deep learning methods like ours assume ample RAM and GPU time. Designing an economical both in terms of memory and computational time is one of our priorities in future directions. We briefly discussed this point in the main text.

Minor comments:

1) Line 7: "[…] by taking the difference between the spliced and unspliced counts[…]" I think the

word “difference” is misleading, as it points towards a simple mathematical subtraction, which is not what RNA velocity tools consider.

We agree with the reviewer. It is about measuring the rate parameter in ordinary differential equations. We revised the text as: “the divergence of the spliced counts from the unspliced” to be more general. Additionally, we added: “More precisely, having the two types of mRNA counts, we can solve for ordinary differential equations of transcriptional dynamics and estimate the splicing and mRNA decay rate parameters.”

2) Line 53: Q/C not defined.

We apologize for the oversight. We revised it as “quality control” instead of “Q/C.”

3) Lines 84-85 mention dropout, however this used to be a concern in early single-cell RNAseq technologies. It is generally believed not to be an issue in droplet single-cell RNA-seq technologies.

We clarified the text and made a distinction between a drop-out event due to technology and low intrinsic counts within a single cell.

4) Lines 86-89:“For instance […] values (Fig. 1B on the right).” The Authors present the 3 cases in Figure 1B as examples of “sparsity due to technologies”. However, I disagree with this view: spliced and unspliced abundances can differ; this does not imply a dropout event.

Again, we rewrote the paragraph removing unvetted claims.

5) Related to the above comment, in the caption of Figure 1, the Authors argue that: “The red dashed line indicates steady state, where the spliced and unspliced gene is in the same amount.” To me, this is inaccurate: spliced and unspliced abundances can also differ significantly at steady state (e.g., depending on the degradation rates).

Yes, we agree with the comment. The red dashed lines were not meant to indicate steady states. Our intention was to help readers grasp systemic differences for one gene. The legend texts were erroneously written as a result of miscommunication between the authors. We fixed it.

6) Methods: “Dieng and coworkers” I think should be followed by (year) and [citation].
We thank the reviewer for pointing out the oversight. We revised the paragraph.

7) Methods: “(1) We introduces a Bayesian hierarchical prior […]” Point (2) is on a separate line (bullet point style), so I think (1) also should.

Yes, we changed it as suggested. Thank you.

8) Methods: “with respect the variational parameters”.I think that “to” is missing after “with respect”.

Yes, we changed it as suggested. Thank you, again.

Reviewer #2

The authors use a deep variational topic model to learn the common and additional latent space for

spliced and unspliced transcripotomic data. The model is mathematically sound and the findings are pretty interesting. I only have a few minor comments.

We are deeply grateful to the reviewer for reading our manuscript and sharing insights with us. Since we will continue to pursue future research in this direction, we found the reviewer's comments were very helpful in many ways.

1. It would be nice to compare the BALSAM method to other types of latent space decomposition method like NMF. I know one is a probabilistic model and another is a matrix factorization. I am curious about what's the advantage of the probabilistic one. I think computational wise, NMF should be much faster.

Thank you for your suggestions. We conducted benchmark analyses based on realistic simulations. As suggested, we compared DeltaTopic and BASALM models with PCA, NMF, and LIGER (a variant of NMF). We wrote a new paragraph to briefly discuss the benefit of using sparse probabilistic approach.

2. I am curious about the broad application of the BALSAM method, so it would be helpful to try it on some generic transcriptomic data and see how the latent space it produced can compare to the others in terms of dimension reduction and clustering performance etc.

Our benchmark results also demonstrate that BALSAM outperforms traditional latent factor methods, such as PCA and NMF. Although more extensive benchmark studies across many different data sets will be of research interest, we regret that the variational inference algorithm used in BALSAM is not necessarily ideal for extensive benchmark analysis across many data sets. Since we are currently developing a scalable algorithm tailored for sparse topic models, we decided to leave more extensive simulation studies for future direction.

3. It would be helpful if the running time of the deltatopic method can be reported.

We added a new paragraph to report run time, GPU and memory usage at the end of Result section. Thank you for your suggestions.

4. Some typos: line 71 gell-topic -> cell-topic; appendix: namely and xi -> namely and □.

We fixed those typos. Thank you.

Referees' report, second round of review

Reviewer #1: The Authors have successfully and fully addressed all my comments, particularly regarding software availability (nice online interface too).
I am satisfied with their new submission and suggest the paper to be accepted as it is.

Kind regards,
Simone Tiberi
The University of Bologna

Reviewer #2: Comments enter in this field will be shared with the author; your identity will remain

anonymous.

My concerns have been addressed successfully, congrats on an improve paper.

Authors' response to the second round of review