Rebuttal letter for PeerJ

In the letter below, the reviewer and editor comments are in plain text, and our response is *in italics*.

Editor's comments

`MINOR REVISIONS`

Both reviewers have a few comments that should be straightforward to address.

Response: Thank you for the careful reviews of the paper, and sorry for the long delay in submitting the revision. The delay was due to other intervening non-scientific factors.

Reviewer 1 (Eric Brenner) comments

Basic reporting

- On line 99, "results" should be singular: "result". Other than that, the writing was easy to read, and I did not notice any other grammar or spelling mistakes.

Response: Thanks for catching this typo, it is has been corrected.

- Paragraph starting on line 37. Another strong argument for using unequal proportions of cell types when assessing the accuracy of a scRNA-seq technique is that it is more representative of what you'd find inside of an organism. When working with blood or a tissue containing multiple cell types, it's unlikely that those cell types will exists in equal proportions. Make sure to emphasize this at some point.

Response

- At some point in the Introduction is important to make it clear that other people have intentionally used unequal proportions of cell types in QC analysis of an scRNA-seq technique. For example, see Figure 2F of Rosenberg et al, 2017. However, your work is novel in that it provides a detailed explanation of how to calculate the multiplet frequency in these scenarios.

Response

- Lines 44-47. Give some brief detail about the Poisson distribution for readers who may not be familiar with it.
- A figure with diagrams showing how the scRNA-seq techniques work may be helpful to some readers.
- Instead of calling R code from within Python (which only works for Python 2 and not Python 3, as far as I know) it would be best to just provide two scripts for readers to choose from that each contain the whole analysis pipeline with one script written purely in Python (no R code) and another purely written in R.

Experimental design

- See my previous comment regarding the Rosenberg et al paper.

- How closely do you expect the number of cells per droplet to follow a Poisson distribution? From my understanding of microfluidics devices, there would be some upper limit to how many cells could physically fit into a single droplet. This concern may be negligible statistically, but it should still be addressed.

Validity of the findings

no comment

Comments for the author

no comment

Reviewer 2 (Peter Sim) comments

Basic reporting

The article is well-structured and clearly written with appropriate references and easy-to-follow code (including an html rendering of a Python Jupyter Notebook) for recapitulating the tables in the paper.

Experimental design

The main purpose of this article was to derive a general equation for evaluating the multiplet frequency in a single-cell RNA-Seq experiment (e.g. from a mixed species experiment). Multiplets represent a major problem in single-cell RNA-Seq (and single cell analysis in general), as they often give profiles that resemble hybrid cell types or states, when they actually originate from a spurious mixture of two or more different cells in a single chamber. The author clearly articulates the problem that the current literature only addresses the case that two species are mixed in equal proportion. He provides an important experimental scenario in which this problem is highly relevant – namely the case that an experimenter includes a small number of cells from a different species as a "spike-in" sample during single-cell RNA-Seq for internal evaluation of the multiplet frequency. He then describes a detailed and straightforward derivation, based on simple Poisson statistics, of an equation that accommodates the more general case that the two species are mixed in arbitrary proportions. The derivation is mathematically sound and provides a straightforward way to evaluate multiplet frequencies from experimentally accessible observables. The methodology used to compute example calculations is easy to replicate, because thoroughly commented code is provided.

Validity of the findings

The findings presented here are valid and useful. The conclusions are well-supported by a straightforward mathematical derivation – an analytical solution to the proposed problem is provided.

Comments for the author

I have a few questions and comments regarding the potential for more general application of the analysis described here:

1) In many single-cell RNA-Seq experiments, unsupervised clustering reveals very discrete cell types that are readily distinguishable, but from the same species. The author focuses on the case that two cells from two different species are mixed in arbitrary proportion, but could these results be extended to the case that an arbitrary number of cell types is mixed in varying proportions? As a concrete example, if I am looking at a blood sample, I might expect to readily separate monocytes, B cells, T cells, etc. Could the framework described here be used to calculate an expectation value of, for example, the doublet frequency of each pair of cell types? This may be beyond the scope of the study described here, but may be worth a comment in the manuscript.

2) Throughout the paper, the author focuses on what I would call "statistical multiplets". However, in many single-cell RNA-Seq experiments, this is not the only source of multiplets (and in some cases, not even the dominant source). For example, incomplete cellular dissociation could result in sequencing of multiplets simply because the cells were stuck together in the original tissue and remain stuck together during the profiling experiment. In my own work, I have noticed an enrichment in apparent multiplets coming from cell types that I know are interacting with each other in the tissue I am profiling. This makes the application described in 1) particularly important. Having an expectation value for the "statistical multiplet" frequency could provide a framework for evaluating whether or not the observed multiplet frequency for a pair of cell types is higher than one would expect, thereby implying incomplete dissociation. Again, this may be beyond the scope of the study described here, and I leave this to the author's judgement, but I think it would be worth commenting on the distinction between "statistical multiplets" and multiplets arising from other sources.