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: This is a good point. I have added a sentence explaining that in naturally occurring mixtures of cells of different types (such as blood or tissue), the different cell types will usually be present at unequal proportions.

- 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: This is a good point. I have added a reference to the Rosenberg et al study in the Introduciton.

- Lines 44-47. Give some brief detail about the Poisson distribution for readers who may not be familiar with it.

Response: I have added text explaining that the Poisson distribution is accurate when cells are loaded into droplets randomly and independently.

- A figure with diagrams showing how the scRNA-seq techniques work may be helpful to some readers.

Response: Given the wealth of different scRNA-seq techniques that can lead to Poisson loading, I have decided against trying to illustrate them graphically. I agree that such diagrams are useful, but they are present in most of the references cited in the first paragraph of the Introduction. My expectation is that this paper will be of interest primarily to experts who are already familiar with the basics of scRNA-seq and/or have read some of the references that explain these methods.

- 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.

Response: This is a good idea. I now provide two separate Juypter notebooks, one that performs the calculations purely in R, and one that performs the calculations purely in Python.

Experimental design

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

Response: This point has been addressed; see above.

- 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.

Response: This is an excellent point. I am not able to directly quantify how "Poisson" different methods actually are from available data, so I have added several sentences to the Discussion that discuss how factors such as cell clumping could cause deviations from Poisson behavior, and how this might affect the calculations.

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.

Response: Thanks for the nice summary.

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.

Response: Thanks for the nice summary.

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.

Response: This is an interesting idea. My concern would be that in the case of cell-type clustering, the multiplets might affect the clustering and subsequently the cell-type counting. In the case of mixing human and mouse cells, the classification is unambiguous in terms of single-species or mixed-species. But in cell type classification, if there are a substantial number of multiplets, they might skew the clustering, thereby changing the numbers of cells of each type that go into the calculation. I am not sure how big a problem this would be in practice, but because of this concern I am reluctant to suggest the approach be applied in such a scenario without extensive further investigation. For this reason, I have simply included sentences to the Discussion that emphasize that the method is restricted to cases where the cell types have already been classified.

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.

Response: This is an excellent point, and similar to one made by Reviewer 1 about the validity of the Poisson assumption. I have added text to the Discussion (second paragraph) emphasizing how the calculation assumes a Poisson distribution which could be violated by factors such as cell clumping. However, I am not certain how to extend the calculations to identify deviations from Poisson behavior with the input data that the approach currently assumes. I think that an additional data point (such as the number of empty droplets) would need to be available in order to determine if the multiplet frequency is actually higher than expected under Poisson. We can say that non-Poisson behavior should make the calculations here conservative provided that there is not a bias towards clumping of cells of the same type; this fact is now noted in the Discussion.