Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments

Jesse D. Bloom[1]

[1]**Fred Hutch Cancer Research Center, Seattle, WA 98109**

Corresponding author:
Jesse D. Bloom[1]

Email address: jbloom@fredhutch.org

ABSTRACT

In single-cell RNA-sequencing, it is important to know the frequency with which the sequenced transcriptomes actually derive from multiple cells. A common method to estimate this frequency is to mix two different types of cells (e.g., human and mouse), and then determine how often the transcriptomes contain a mix of transcripts from both cell types. When the two cell types are mixed in equal proportion, the calculation of the multiplet frequency from the frequency of mixed transcriptomes is straightforward. However, there are no published descriptions of how to calculate the multiplet frequency in the general case when the cell types are mixed unequally. Here I derive equations to analytically calculate the multiplet frequency from the numbers of observed pure and mixed transcriptomes when two cell types are mixed in arbitrary proportions, under the assumption that the loading of cells into droplets or wells is Poisson.

INTRODUCTION

Many methods for single-cell RNA sequencing involve partitioning cells into barcoded droplets (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017), wells (Gierahn et al., 2017), or combinations of wells (Cao et al., 2017). The fundamental principle of these methods is that the number of possible barcoded partitions exceeds the number of cells. Therefore, most partitions contain at most one cell if the number of cells per partition is Poisson distributed. However, some fraction of the partitions contain multiple cells, and estimating this *multiplet frequency* is an important aspect of experimental quality control. Crucially, for many methods, the actual number of possible partitions is not precisely known, and all that is observed is the number of different barcoded transcriptomes after sequencing (i.e., the number of non-empty partitions).

The most common method to determine the multiplet frequency is to perform the experiment on a mix two types of cells (e.g., human and mouse). During the analysis of the sequencing results, each partition can be classified as containing exclusively transcripts from one of the two cell types, or a mix of transcripts from both cell types. Partitions that contain a mix of transcripts must be multiplets. If the two cell types are mixed in equal proportions and the average number of cells per partition is low (so that most multiplets are doublets), then the multiplet frequency can be estimated as simply twice the fraction of non-empty partitions that contain a mix of both cell types. The logic is that all the multiplets are doublets, and only half the doublets will have cells of both types (the others will have two cells of the same type). This approach has been used to estimate the multiplet rate during the initial prototyping of most single-cell RNA sequencing methods (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017; Gierahn et al., 2017; Cao et al., 2017).

However, in some cases the two cell types may be mixed in unequal proportions. Unequal mixing could arise simply from error during cell counting, but it can also be a desirable aspect of experimental design. For instance, if the researcher is actually interested in the human cells and simply wants to include an internal control to estimate the multiplet frequency during each new experiment, then (s)he may want to add substantially smaller numbers of mouse cells to estimate the multiplet frequency while still mostly obtaining data for human cells. But when the cells are mixed unequally, it is no longer valid to estimate the multiplet frequency as simply twice the fraction of non-empty partitions that contain a mix of both cell types. Surprisingly, I could find no published descriptions of the correct way to calculate the multiplet frequency from unequal mixes of two cell types. Here I remedy this gap in the literature by deriving the equations to compute the multiplet frequency when the cells are mixed in arbitrary proportions under the assumption that the number of cells per partition is Poisson distributed.

RESULTS AND DISCUSSION

Derivation of multiplet frequency from observed numbers of pure and mixed-cell droplets

We consider the case in which cells of two types (e.g., human and mouse) are distributed into individual barcoded droplets, although the same logic applies if the cells are distributed into barcoded wells or combinations of wells. We assume the sequencing data have been analyzed so that each non-empty droplet can be classified as containing at least one cell of type 1, at least one cell of type 2, or cells of both types. We will refer to the number of droplets in each of these three groupings as N_1, N_2, and $N_{1,2}$, respectively. For instance, the 10X `cellranger` pipeline (version 2.1.1) returns these numbers as the "Estimated Number of Cell Partitions."

The only assumption of our analysis is that the number of cells per droplet is Poisson distributed. Let μ_1 be the average number of cells of type 1 per droplet, and μ_2 be the average number of cells of type 2 per droplet. The average number of cells of any type per droplet is then $\mu_1 + \mu_2$. So the probability that a droplet contains at least one cell of any type is

$$\begin{aligned} \Pr(c \geq 1) &= 1 - \Pr(c = 0) \\ &= 1 - e^{-\mu_1 - \mu_2}. \end{aligned} \tag{1}$$

Likewise, the probability that a droplet contains multiple cells of any type (e.g., a multiplet) is

$$\begin{aligned} \Pr(c \geq 2) &= 1 - \Pr(c = 0) - \Pr(c = 1) \\ &= 1 - e^{-\mu_1 - \mu_2} - (\mu_1 + \mu_2)\, e^{-\mu_1 + \mu_2}. \end{aligned} \tag{2}$$

The multiplet frequency M is simply the probability that a droplet with at least one cell actually contains multiple cells, which is

$$\begin{aligned} M &= \frac{\Pr(c \geq 2)}{\Pr(c \geq 1)} \\ &= 1 - \frac{(\mu_1 + \mu_2)\, e^{-\mu_1 + \mu_2}}{1 - e^{-\mu_1 - \mu_2}}. \end{aligned} \tag{3}$$

However, to evaluate this expression for M, we need to know the values of μ_1 and μ_2.

We can write down equations for μ_1 and μ_2 by again using the fact that the number of cells per droplet is Poisson distributed. Specifically, if N is the total number of droplets (empty and non-empty), then the expected number of droplets that have at least one cell of type 1 is $N \times \Pr(c_1 \geq 1) = N\left(1 - e^{-\mu_1}\right)$. The observed number of droplets with at least one cell of type 1 is N_1, so setting the observed number equal to the expected number gives us an equation for μ_1,

$$N_1 = N\left(1 - e^{-\mu_1}\right). \tag{4}$$

We can easily solve this equation for μ_1 to yield

$$\mu_1 = -\ln\left(\frac{N - N_1}{N}\right), \tag{5}$$

and likewise for μ_2 we have

$$\mu_2 = -\ln\left(\frac{N - N_2}{N}\right). \tag{6}$$

Equations 5 and 6 give us a way to determine the values (μ_1 and μ_2) needed to calculate the multiplet frequency (Equation 3) in terms of the experimental observables N_1 and N_2. Unfortunately, these two equations also require knowledge of the total (empty and non-empty) number of droplets N, which is not directly observable from the sequencing data.

However, we can take advantage of another relationship to calculate N. The fraction of all (empty and non-empty) droplets that contain cells of both type is $\frac{N_{1,2}}{N}$, and we expect this fraction to simply be the product of the probability that a droplet contains at least one cell of type 1 with the probability that a droplet contains at least one cell of type 2, which in mathematical terms can be stated as $\Pr(c_1 \geq 1 \wedge c_2 \geq 1) = \Pr(c_1 \geq 1) \times \Pr(c_2 \geq 1)$. Therefore, we have

$$\frac{N_{1,2}}{N} = \frac{N_1}{N} \times \frac{N_2}{N}. \tag{7}$$

This equation can be solved to give

$$N = \frac{N_1 N_2}{N_{1,2}}, \tag{8}$$

which can be completely evaluated in terms of the experimental observables. So we can use Equations 5, 6, and 8 to calculate μ_1 and μ_2 in terms of the experimental observables, and then use those results to calculate the multiplet frequency via Equation 3. Therefore, we now have an analytic solution to the multiplet frequency in terms of the three experimental observables.

Implementation and example calculations

https://github.com/jbloomlab/multiplet_freq/blob/master/calcmultiplet.ipynb

experiment	human droplets	mouse droplets	nonempty droplets	human and mouse droplets	multiplet freq	twice cross celltype freq
1	2005	2005	4000	10	0.005	0.005
2	2050	2050	4000	100	0.049	0.050
3	2500	2500	4000	1000	0.425	0.500

Table 1. Caption

experiment	human droplets	mouse droplets	nonempty droplets	human and mouse droplets	multiplet freq
1	2050	2050	4000	100	0.049
2	3050	1050	4000	100	0.065
3	3550	550	4000	100	0.110
4	3850	250	4000	100	0.245
5	3950	150	4000	100	0.459

Table 2. Caption

CONCLUSIONS

METHODS

The LaTex source for this paper, the Jupyter notebook that implements the calculations, and all materials associated with the writing and review of the paper are publicly available in a GitHub repository at https://github.com/jbloomlab/multiplet_freq.

REFERENCES

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667.

Gierahn, T. M., Wadsworth II, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C., and Shalek, A. K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14(4):395.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049.

Supplemental file 1. A Jupyter notebook that implements the calculations in Python and R functions, and does the calculations for the examples shown in the tables in this paper.

Supplemental file 2. This file contains an HTML rendering of the Jupyter notebook in Supplemental file 1.