# 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 at which the sequenced transcriptomes actually derive from multiple cells. A common method to estimate this multiplet frequency is to mix two different types of cells (e.g., human and mouse), and then determine how often the transcriptomes contain 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. But surprisingly, 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). As long as the number of possible partitions exceeds the number of cells, then most partitions will contain at most one cell. However, some fraction of the non-empty partitions will contain multiple cells, and estimating this *multiplet frequency* is an important aspect of experimental quality control.

The most common method to determine the multiplet frequency is to mix two types of cells (e.g., human and mouse). During the analysis of the sequencing results, each non-empty partition can be identified as containing transcripts from one or both of the two cell types. Partitions that contain a substantial number of transcripts from both cell types must be multiplets. If the two cell types are mixed equally 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 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 frequency during the 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, or it could be an intentional aspect of experimental design (Rosenberg et al., 2018). 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 fewer mouse cells so that most of the resulting data is for the human cells. In addition, when analyzing naturally occurring mixtures of cells of multiple types, the different cell types will usually be present in unequal proportions. 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 how 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. This Poisson assumption is accurate when cells are loaded randomly and independently into partitions.

## METHODS

The LaTex source for this paper, the Jupyter notebooks that implement 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`. The Jupyter notebooks are also available in Supplemental file 1 and 3, and HTML renderings of the notebooks are in Supplemental file 2 and 4.

## RESULTS

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

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. 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. I 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 the derivation is that the number of cells per droplet is Poisson distributed. Let $\mu_1$ be the average number of cells of type 1 per droplet, and $\mu_2$ 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, evaluating this expression for $M$ requires the values of $\mu_1$ and $\mu_2$.

We can write down equations for $\mu_1$ and $\mu_2$ 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 $N \times \Pr(c_1 \geq 1) = N(1 - e^{-\mu_1})$. 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 $\mu_1$,

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

This equation is easily solved for $\mu_1$ to yield

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

and likewise for $\mu_2$,

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

Equations 5 and 6 give us a way to determine the values ($\mu_1$ and $\mu_2$) 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 types is $\frac{N_{1,2}}{N}$, and this fraction is simply 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,

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

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. Equations 5, 6, and 8 can be used to calculate $\mu_1$ and $\mu_2$ in terms of the experimental observables, and those results used to calculate the multiplet frequency via Equation 3. This provides an analytic solution for the multiplet frequency in terms of the three experimental observables.

## Implementation and example calculations

A simple function to perform the calculations described in the previous subsection is implemented in Python in the Jupyter notebook found at `https://github.com/jbloomlab/multiplet_freq/blob/master/calcmultiplet.ipynb`, and in R in the Jupyter notebook found at `https://github.com/jbloomlab/multiplet_freq/blob/master/calcmultiplet_R.ipynb` (see also Supplemental files 1, 2, 3, and 4). To illustrate the calculations, I used this function to calculate the multiplet frequency for hypothetical data.

First, consider hypothetical data in which the two types of cells are mixed in equal proportions. Prior papers have approximated the multiplet frequency from such experiments as simply twice the fraction of non-empty droplets that contain cells of both types (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017; Cao et al., 2017), which is $\frac{N_{1,2}}{N_1+N_2-N_{1,2}}$ in the notation defined in the previous subsection. Table 1 shows that the exact equation derived in the previous subsection gives very similar results to this approximate method as long as the multiplet frequency is low. When the multiplet frequency becomes high, the approximate method starts to overestimate the true multiplet frequency, since it fails to account for the fact that some multiplets will contain more than two cells.

Next, consider hypothetical data in which the two types of cells are mixed in unequal proportions. Table 2 shows the multiplet frequencies for several such experiments. An interesting aspect of the results is that at high multiplet frequencies and very unequal cell proportions, the multiplet frequency is

| 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.** Multiplet frequencies for three hypothetical experiments in which human and mouse cells are mixed equally. The multiplet frequencies calculated using the exact method described here (column *multiplet freq*) are very similar to those obtained simply by multiplying by two the fraction of non-empty droplets that contain cells of both types (column *twice cross celltype freq*). However, the two methods are slightly different at higher multiplet frequencies, since the latter method fails to account for multiplets that have more than two cells.

| 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.** Multiplet frequencies for five hypothetical experiments in which human and mouse cells are mixed unequally.

substantially *lower* than the fraction of droplets containing the rarer cell type that contain a mix of both cell types. The reason is that multiplets (particularly higher-order ones) become more and more likely to contain at least one cell of the rarer type relative to droplets that contain only one cell. For instance, in the final experiment in Table 2, two thirds of the droplets containing mouse cells have a mix of both cell types, yet less than half the non-empty droplets are multiplets (the multiplet frequency is 0.459). This somewhat non-intuitive result illustrates the importance of using the correct mathematical relationship to calculate the multiplet frequency when cell types are mixed unequally.

## CONCLUSIONS

I have described how to calculate the multiplet frequency in single-cell RNA sequencing experiments in which two cell types are mixed in arbitrary proportions. It is important to note that this calculation requires that the sequencing data have already been analyzed to determine whether each partition contains a non-negligible number of transcripts from each cell type, but many common analysis programs (such as the 10X `cellranger` pipeline) already do this.

The calculation also assumes that the number of cells per droplets follows a Poisson distribution. While many single-cell RNA sequencing methods are designed to partition cells in a way that concords with this assumption (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017; Gierahn et al., 2017; Cao et al., 2017), it is possible that cell clumping or other factors could bias certain partitions to contain more cells than expected under a Poisson distribution. In such a scenario, the calculations in this paper would overestimate the true multiplet frequency if the clumping is equally likely across cell types, but could underestimate the true multiplet frequency if intra-cell-type clumping is more likely than inter-cell-type clumping.

Finally, the approach in this paper only calculates the multiplet frequency—it does *not* actually identify the multiplets so that they can be removed from downstream analyses. For that purpose, other more sophisticated approaches have been developd (Ilicic et al., 2016; Stoeckius et al., 2017; Kang et al., 2018; Wolock et al., 2018; DePasquale et al., 2018). Nonetheless, simply calculating the multiplet frequency from the data returned by standard pipelines such as the 10X `cellranger` is important for many purposes, and the results here enable that to be done regardless of the proportions at which the cell types are mixed.

## 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., , Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667.

DePasquale, E. A., Schnell, D. J., Valiente, I., Blaxall, B. C., Grimes, H. L., Singh, H., and Salomonis, N. (2018). Doubletdecon: Cell-state aware removal of single-cell rna-seq doublets. *bioRxiv*, page 364810.

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.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1):29.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., Gate, R. E., Mostafavi, S., Marson, A., Zaitlin, N., Criswell, L. A., and Ye, C. J. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89.

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., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., and Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B., Smibert, P., and Satija, R. (2017). Cell "hashing" with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *bioRxiv*, page 237693.

Wolock, S. L., Lopez, R., and Klein, A. M. (2018). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *bioRxiv*, page 357368.

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., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schanll-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (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 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.

**Supplemental file 3.** A Jupyter notebook that implements the calculations in R, and does the calculations for the examples shown in the tables in this paper.

**Supplemental file 4.** This file contains an HTML rendering of the Jupyter notebook in Supplemental file 3.