Journal Title Here, 2024, 1–7
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

More cells, more doublets in sample-barcoded single-cell data

George Howitt,[1, 2] Gunjan Dixit,[1, 2] Rotem Aharon,[1,2] Victoria Streeton-Cook,[1] Ling Ling,[3] Peter F. Hickey,[3,4] Daniela Amann-Zalcenstein,[3,4] Liam Gubbels,[5] Shivanthan Shanthikumar,[5,6,7] Sarath Ranganathan,[5,6,7] Melanie Neeland,[5,6] Jovana Maksimovic[1, 2] and Alicia Oshlack[1, 2, 8,*]

[1]Computational Biology Program, Peter MacCallum Cancer Centre, Parkville, VIC, Australia, [2]Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, VIC, Australia, [3]The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia, [4]Department of Medical Biology, University of Melbourne, Parkville, VIC, Australia, [5]Respiratory Diseases, Murdoch Children's Research Institute, Parkville, VIC, Australia, [6]Respiratory and Sleep Medicine, Royal Children's Hospital, Parkville, VIC, Australia, [7]Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia and [8]School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, Australia

[*]alicia.oshlack@petermac.org

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Sample barcoding allows deconvolution of multiplets in multiplexed droplet-based single-cell RNA-sequencing experiments. However, this is only possible when each cell comes from a different sample. As the number of cells in a droplet increases, the probability of two or more cells coming from the same sample increases rapidly. We show that the number of these unresolvable multiplets is greater than previously estimated for the 10X Flex scRNA-seq protocol, and provide a formula for estimating the fraction of multiplets in a data set given a measured average droplet occupancy and number of unique samples in a pool. We also show that existing doublet detection tools should be applied to Flex data to identify these multiplets, and demonstrate that filtering out barcodes identified by these tools improves downstream analysis.

Introduction

As single-cell RNA-sequencing (scRNA-seq) experiments grow in scale, performing separate captures for individual samples becomes prohibitively expensive. Pooling multiple samples in a single experiment both reduces cost and minimises batch effects, but introduces a new problem – how to associate individual cells with their sample of origin. A common method involves tagging the samples with either antibody or lipid-linked unique oligo barcodes prior to pooling. These "hashtag" oligos attach to either proteins [14] or lipids [11] on the cell surface and can be sequenced along with the gene expression, and various methods have been developed to identify the sample of origin from the expression counts of the hashtags [9]. However, hashtag-based multiplexing has several drawbacks. Firstly, tagging introduces an additional step into the sample preparation workflow, and may be more challenging for certain tissue types. Poor staining can cause many cells to be unidentifiable, even when the underlying RNA expression data is high quality [3]. re, especially for poor quality hashtagging [9].

In droplet-based scRNA-seq protocols, cells are encapsulated within droplets which contain oligo barcoded beads that hybridise to each RNA molecule in the cell. This allows for the allocation of reads to droplets, and, assuming each droplet contains either zero or one cell, each read can be associated with a single cell. If droplets contain two or more cells (called doublets, or more generally multiplets), the RNA expression from all the cells is pooled together in the expression counts matrix, creating artifacts which may be erroneously identified as novel or transitional cell types. While multiplets can be identified based on their RNA expression profiles or using hashtags, they must be discarded from any downstream analysis [13]. Therefore, excessive numbers of multiplets are a major concern in droplet-based scRNA-seq experiments, particularly as the main approach for avoiding them is to reduce the capture throughput, thereby producing many more empty than cell-containing droplets, which in turn increases costs.

The Flex protocol from 10X Genomics addresses both of the above issues by using barcoded gene-specific probes that hybridise to individual RNA molecules in the cell. These barcodes can include an oligo that

identifies a specific sample, with available kits containing 4 and 16 unique sample barcodes. The uniquely barcoded gene probes are added to each sample prior to pooling. Probe hybridisation is then followed by encapsulation within droplets containing GEMs (Gel beads-in-emulsion) similar to other 10X droplet-based protocols. After sequencing, each read then includes a barcode from both the sample of origin as well as the droplet, allowing immediate demultiplexing of pooled samples and resolution of multiplets from distinct samples within the same droplet. This removes the need for separate hashtag-staining and demultiplexing steps in the scRNA-seq workflow, and the resolution of multiplets allows for higher throughput in the library preparation stage. Consequently, according to the manufacturer, up to 128,000 cells from up to 16 samples can be run on a single lane of a Chromium X instrument [5].

However, while cells from different samples in the same droplet can be resolved with Flex, cells from the same sample remain unresolvable, and the number of these increases steeply with the occupancy of droplets. This is analogous to the "birthday paradox" in classical probability, where the probability of two people in a room sharing the same birthday increases faster than expected by intuition, as the number of people in the room increases [12]. Here we demonstrate that the number of multiplets found in real Flex data is significantly higher than initially expected, and describe a mathematical model for estimating the number of multiplets that is closer to the observed value.

Results

Flex data sets

We perform our analysis using two data sets acquired from ten 16-plex Flex runs. The first data set, the earlyAIR data set, comes from an ongoing study to create a single-cell atlas of the paediatric airway, and comprises nine 16- plex batches from five tissue types, comprising both solid and liquid tissues from healthy donors. Some donors provided multiple tissues, and the batches contain between 25,486 and 139,926 cells after pre-processing (see Methods). The second data set, the PBMC data set, comprises 16 replicates of peripheral blood mononuclear cells taken from a healthy donor and is publicly available from 10X Genomics (see Data availability). The PBMC data set contains 135,785 cells after pre-processing.

Verification of doublet finders with known multiplets

Tests of doublet-finding methods have to date relied upon a ground truth determined from alternative experiments, such as genotype information [13], hashtag-based demultiplexing [4], DNA barcoding [18], or synthetic data [15]. Flex allows for a novel means of testing doublet finding directly using the RNA sequencing output. In a Flex experiment, due to superloading, many droplets contain multiple cells and cell calling is performed separately for each sample barcode. This means the initial step in a standard Flex analysis is to separate reads based on their combined droplet and sample barcode and then call cell-containing or empty "droplets" based on the data separated by samples. A droplet/sample combination may look empty even when there is a "cell" from a different sample in the same droplet. The resulting counts matrix has cells labeled by both the droplet barcode and sample barcode. Aggregating the counts for each "droplet" across the sample barcodes creates a droplet-level counts matrix analogous to 3' scRNA-seq experiments. Since the number of cells (from unique samples) called by CellRanger in each droplet is known, the droplet-level counts matrix provides a ground truth data set for testing doublet finders: droplets with only one cell are singlets and droplets with two or more cells are doublets. Note, this section is not intended as a comprehensive benchmark of doublet-finding methods, which may be found in [13]. Instead, we apply two of the best-performing packages

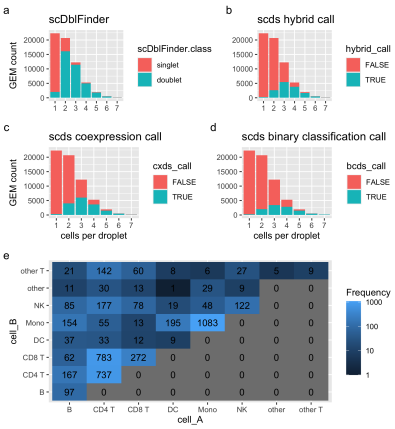

Fig. 1: a)-d) Performance of doublet-finders on droplets with occupancy determined by CellRanger calls. Each bar represents the identified number of cells in a droplet based on the sample deconvolution. The blue portion represents the number of detected doublets from the doublet detection tool. Ideally, a tool will only detect singlets where there is one cell per droplet and doublets when there is more than one cell. e) Cell type composition of droplets containing two cells called by CellRanger which are called singlets by scDblFinder. The majority (¿50%) of cells misclassified as singlets are homotypic doublets.

with low computational overhead recommended in [13]: scds [2] and scDblFinder [6] and evaluate which performs best on our Flex data sets. scds has two methods for doublet detection, one based on marker co-expression and another based on binary classification, as well as a hybrid classifier that combines both methods. We use all approaches and some details of how each method works and the parameters chosen can be found in the Methods section.

Figure 1 shows the performance of scDblFinder and the two scds methods as well as the hybrid classifier on the droplet-aggregated 10X PBMC data set. For this data, scDblFinder (Figure 1a) clearly outperforms the scds methods (Figure 1 b-d), with 90% of the multiply-occupied droplets called as doublets compared to only 25%-40%. 22% of droplets containing two cells are called singlets by scDblFinder. Doublet-finding methods are known to perform poorly on droplets containing multiple cells of the same type (homotypic doublets), as these are unlikely to show discrepant transcription profiles [16, 13]. We use the reference mapping package Azimuth and its included PBMC reference data set to annotate the PBMC data set [8], and examine the composition of the unidentified doublets. Figure 1e) shows that the majority (50.5%) are homotypic doublets, or have more closely related transcriptional profiles such as CD8+ T cells with CD4+ T cells (17%). scDblFinder calls 9.2% of the CellRanger singlets as doublets, compared to ¡ 2% for scds. In the following sections, we examine doublets in more depth and show that the rate of unresolved doublets predicted by our mathematical model is quite close to the value from scDblFinder.

Measuring multiplet rates in Flex data

Having identified scDblFinder [6] as our preferred doublet-finding method for this study, we apply it to all of our data sets. Figure 2 shows the fraction of doublets called by scDblFinder on each Flex run as a function of the number of cells. Also shown in Figure 2 is the predicted doublet fraction from the documentation for the Flex protocol [5] of 0.8% per 1000 cells

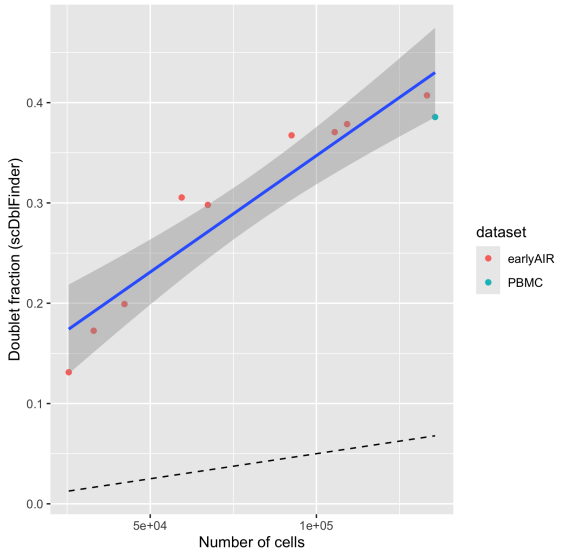

Fig. 2: Doublet fraction (from scDblFinder) vs number of cells called for 10 16-plex Flex data sets. The blue line indicates a linear fit to the data. The black dashed line is the estimated rate from the Flex user guide.

per sample, or 0.05% per 1000 cells for 16-plex runs. A clear linear trend is observed in the results from scDblFinder, with a best fit multiplet rate of 0.23% per 1000 cells, or ≈ 4.6 times more than the estimate from the documentation.

Estimating the resolvable cell fraction

Why is the multiplet rate predicted by doublet finders so much greater than the rate estimated by the manufacturer? Recall that the quoted rate of unresolvable multiplets in Flex comes from dividing the rate for the 3' protocol by the number of samples. This makes sense if each Flex sample is processed separately at the same concentration as an equivalent 3' sample, but the recommended practice is to pool all samples together after probe hybridisation, so that the concentration of cells is up to 16 times greater than a 3' experiment sequencing the same number of cells per sample. 3' experiments are designed to avoid multiplets, so that very few droplets contain more than one cell. In Flex experiments the ability to resolve cells from different samples within the same droplet means that total cell concentrations can be much larger, and droplets containing multiple cells are common (cf. Figure 1). In this section, we perform some calculations showing that increasing the occupancy of droplets leads to a higher number of unresolved cells than previously appreciated.

Consider an ensemble of droplets containing k cells, drawn from a pool containing n samples with equal proportions of cells from all samples. The number of possible configurations of cells within the droplets is the number of combinations of length k drawn from n with repetition allowed, i.e.

$$M_k(n) = \frac{(n+k-1)!}{(n-1)!k!} \quad , \tag{1}$$

where ! denotes the factorial function. Assuming each of these combinations is equally probable, we aim to determine what fraction $F(n,k)$ of cells in droplets containing k cells are "resolvable", which we define as being the only cell from their sample of origin. In general,

$$F(n,k) = \frac{\sum_c f_c}{M_k} \tag{2}$$

where the subscript c denotes a single configuration of cells within a droplet, and f_c is the resolvable fraction of that configuration. Compupting $F(n,k)$ therefore requires determining f_c for each combination.

We define the observed occupancy of a droplet as l, $l \leq k$, as the number of unique samples in a droplet. Droplets with $k = 1$ are all resolvable. For droplets with $k \geq 2$, if all cells in the droplet are from the same sample, i.e. $l = 1$, no cells are resolvable, while if each cell comes from a unique sample, i.e. $l = k$, all cells are resolvable. In general, we can use the concept of integer partition to determine the resolvable fraction of a given combination of cells. A partition is a way of writing an integer k as a sum of integer "parts", e.g. the partitions of 5 are (5), (4 + 1), (3 + 2), (3 + 1 + 1), (2 + 2 + 1), (2 + 1 + 1 + 1), (1 + 1 + 1 + 1 + 1) [1]. In this framework, each part of the partition is a unique sample, the number of parts is the observed occupancy l, and the number of 1s in the partition is the number of resolvable cells, e.g., the partition (2 + 2 + 1) represents a droplet with $k = 5$ consisting of 2 cells from sample A, 2 cells from sample B and 1 cell from sample C. We can define the fraction of resolvable cells in a droplet f in two ways: first, the number of resolvable cells from unique samples in the droplet divided by the total number of cells in the droplet, i.e. the number of 1s in the partition divided by k; second, the number of resolvable cells divided by the number of unique samples in the droplet, i.e. the number of 1s in the partition divided by l. Since we are interested in comparing to experimental data, we use the second definition. For the partitions of 5 written above, this gives $f = 0, 0.5.0, 2/3, 1/3, 0.75, 1$. Since any configuration of cells within a droplet can be represented as a partition, and the resolvable fraction of each partition is easily calculable [1], all that we need to solve (2) is the proportion each partition contributes to the total number of combinations M_k.

To begin, let's again consider the partition (2 + 2 + 1) representing 2 cells from sample A, 2 cells from sample B and 1 cell from sample C. There are n possible choices for sample A. If sample B is the same as sample A, then the droplet instead corresponds to the partition (4 + 1), so there are only $n - 1$ choices for sample B, $n - 2$ choices for sample C, and $n(n-1)(n-2)$ possible configurations for this partition.

In general, we denote each partition of k $P_i(k)$. Define $m_j(P_i)$ the number of occurences of $j \in \{1, \ldots, k\}$ in P_i. To determine the number of combinations corresponding to P_i, $C_i(n,k)$, consider the first non-zero m_j. There are $\binom{n}{m_j}$ possible combinations of samples for these cells, where

$$\binom{n}{m_j} = \frac{n!}{(n-m_j)!m_j!} \tag{3}$$

is the binomial coefficient. For the next non-zero coefficient, m_h, there are only $n - m_j$ unique samples left to choose from, so the number of possible combinations is $\binom{n-m_j}{m_h}$. In general, we expect

$$C_i(n,k) = \prod_{j=1}^{k} \binom{n-d_j}{m_j} \quad . \tag{4}$$

where

$$d_j = \sum_{h=1}^{j-1} m_h \quad , \tag{5}$$

with $d_0 = d_1 = 0$, is a variable that counts how many unique samples have already been chosen. As a check on (4), we find that

$$\sum_i C_i(n,k) = M_k(n) \tag{6}$$

as required. Equation (2) then becomes

$$F(n,k) = \frac{\sum_i f(P_i) C_i}{M_k} \, , \tag{7}$$

where the observed occupancy, $l(P_i)$ is the length of the partition and the resolvable fraction of P_i, $f(P_i) = m_1(P_i)/l$. We show F versus k for

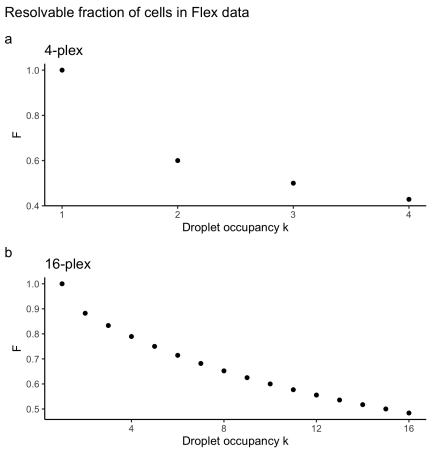

Fig. 3: Fraction of resolvable cells $F(n,k)$ vs occupancy k for (a) $n = 4$ and (n) $n = 16$.

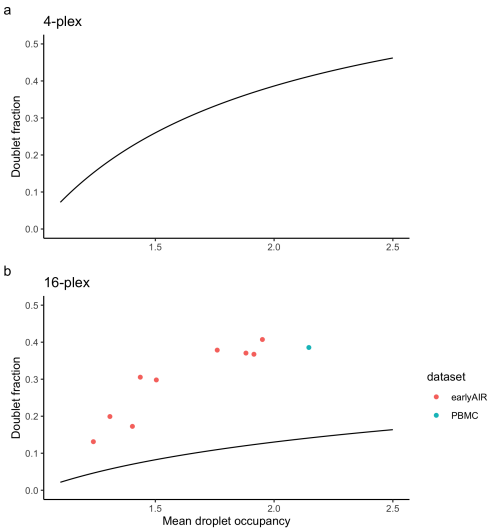

Fig. 4: Calculated fraction of multiplets in Flex data vs mean droplet occupancy given the number of multiplexed samples a) $n = 4$ and b) $n = 16$. Coloured points indicate doublet fraction determined from scDblFinder and observed mean droplet occupancy from earlyAIR and PBMC data sets.

$n = 4$ and $n = 16$ in Figure 3. Figure 3 shows that F decreases steeply with k, with $F(n = 4) < 0.5$, for $k > 2$ and $F(n = 16) < 0.75$ for $k > 5$.

To determine the size of the effect in real data, we need to combine equation (7) with the distribution of droplet occupancy k. A good model for the distribution of observed occupancy l in Flex data is the zero-truncated Poisson distribution, as zeros, or empty droplets, are not included in the CellRanger output (Supplementary Figure 1)

$$p(l) = \frac{\lambda^l}{l!(e^\lambda - 1)}, \tag{8}$$

where λ is a shape parameter related to the mean droplet occupancy $\langle l \rangle$ through

$$\langle l \rangle = \frac{\lambda}{1 - e^{-\lambda}}. \tag{9}$$

We now make a simplifying assumption that the distribution of the true occupancy k can be approximated by the distribution of the observed occupancy l, i.e.

$$p(k) \cong p(l). \tag{10}$$

We make this assumption in order to simplify what follows, and because determination of a relationship between k and l is a non-trivial task requiring a more robust ground-truth data set for known doublets in Flex data than is currently available. The number of cells in droplets with observed occupancy k is then

$$N_{\text{cells}}(k) = N_{\text{droplets}} k p(k). \tag{11}$$

Combining equations (7) – (11), we get the fraction of multiplets in an experiment

$$M_f(n) = 1 - \sum_k \frac{k p(k) F_r(n,k)}{\langle l \rangle} \quad . \tag{12}$$

We show the expected doublet fraction in data sets for $n = 4$ and $n = 16$ as a function of mean observed droplet occupancy $\langle l \rangle$ in Figure 4. Figure 4 also includes the doublet fraction from scDblFinder for the earlyAIR and PBMC data sets in the 16-plex case [6]. Figure 4 shows that the predicted multiplet fraction from equation (12) still underestimates the doublet fraction we observe in real-world data, though by significantly less than the prediction from the Flex User Guide [5]. We propose two potential reasons for this discrepancy. The first is the assumption that the true occupancy distribution is the same as the observed distribution, equation 10. The presence of unresolvable cells implies that the true occupancy is always greater than that observed, see Supplementary Figure 3. Properly accounting for this effect would shift the data points in Figure 4 to the right, bringing them closer to the predicted curve. Another reason is the assumption that all samples have equal numbers of cells. In any experiment, some samples contain more cells than others. We expect that this increases the likelihood of same-sample multiplets, though exactly how is beyond the scope of this work.

Even with the limiting assumptions of our model, however, the fraction of unresolved multiplets in an experiment with a mean droplet occupancy of 2.5 (a typical value for an experiment with the maximum recommended loading of 128,000 cells) is 0.17, almost 3 times greater than the value of 0.064 estimated by 10X. These results demonstrate the necessity of removing doublets from Flex data.

Finally, in Figure 1b we observe that scDblFinder calls 2060 of the droplets in the PBMC data set identified as singlets by CellRanger as doublets. We can estimate the number of observed singlets that are actually doublets using equation (7) and the observed distribution of droplet occupancies. From this, we estimate that ~ 2700 droplets with occupancy $k > 1$ will appear as singlets in this data set. Most of these will have $k = 2$ (2 cells, 2400 droplets), but some droplets can contain more cells from the same sample (Supplementary Figure 2). The number predicted by the model is similar to the value of 2,060 from scDblFinder, with the discrepancy likely due to the difficulty of scDblFinder in identifying homotypic doublets, shown in Figures 1b) and e).

Doublet-finding identifies spurious clusters

In order to demonstrate the effect of doublets on downstream analysis, we perform unsupervised clustering on the PBMC data set at a range of resolutions from 0.1 to 1.0 using the `FindMarkers` function with the SLM algorithm in Seurat. We use two versions of the data: (1) all cells that pass our initial QC filtering, and (2) after removing all of the doublets identified

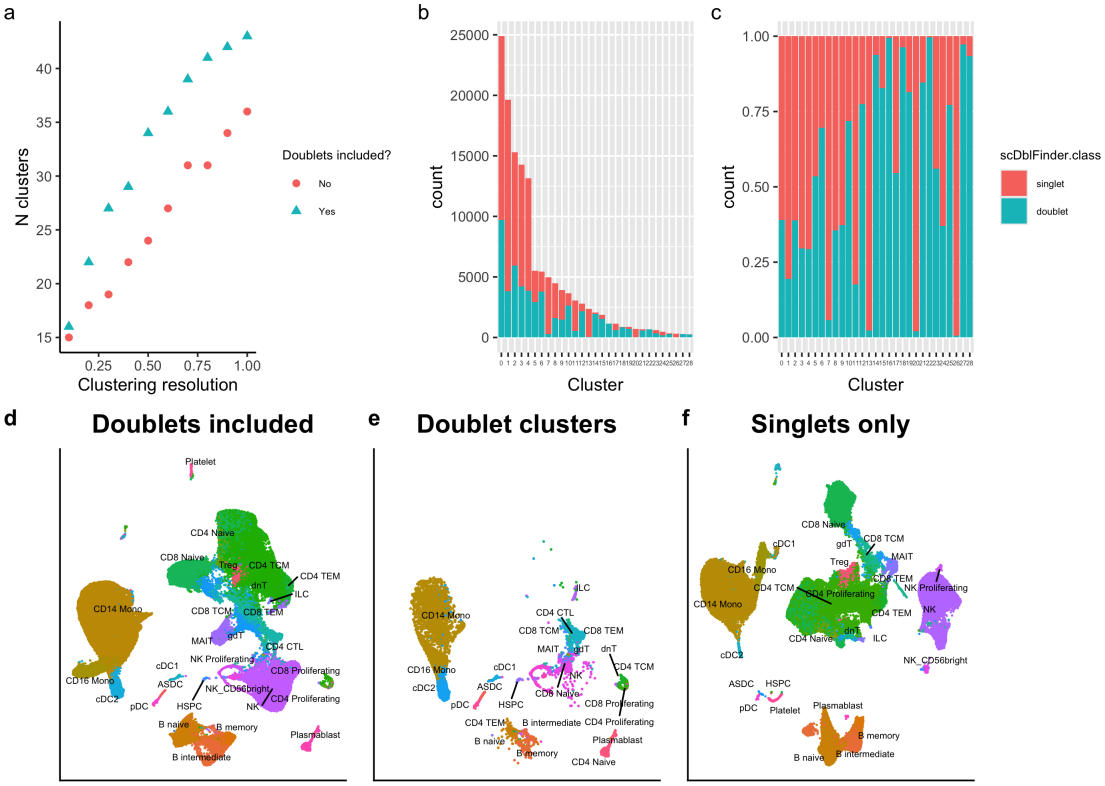

Fig. 5: a) Number of clusters as a function of cluster resolution for PBMC data set before (blue triangles) and after (red circles) removal of doublets identified by scDblFinder. b) Number of cells in each cluster in the PBMC data set at resolution 0.4, coloured by fraction of doublets. c) same as b) but with proportions instead of total counts. d) UMAP showing cells in the PBMC data set with labels from Azimuth. e) Same as d) but showing only cells from clusters with ≥ 75% doublets. f) UMAP with Azimuth labels computed after removing doublets.

by scDblFinder. At every resolution, removing the doublets reduces the number of clusters (Figure 5a), indicating that doublets introduce structure in the data.

We next look at the doublet composition of the identified cluster at resolution 0.4, (Figures 5b and c). We identify 11 clusters that contain more than twice the average doublet fraction of the entire data set. We show the position of all cells in the data set prior to doublet removal, las annotated by Azimuth, on a Uniform Manifold Approximation and Projection (UMAP) plot in Figure 5d. Figure 5e shows only the cells in the 11 clusters identified as being composed of mostly doublets. Figures 5d and e show that these doublet-dominated clusters appear in the regions of the UMAP that bridge multiple cell types, such as CD14 and CD16 monocytes, or T cells and NK cells, lending further evidence for their identity as doublets. Version (1) of the PBMC data contains 52,336 more cells than version (2), which may be responsible for the greater number of clusters. To check this, we randomly remove an equivalent number of cells as the number of doublets from version (1) , so that the resulting data set has the same number of cells as (2) with the same singlet/doublet ratio as (1). After repeating this process 10 times we find that in general, the downsampled data sets contain fewer clusters than (1) but more than (2), and that a significant proportion of clusters are dominated by doublets (see Supplementary Figure 3).

Lastly, Figure 5f shows the UMAP with Azimuth labels after removing the doublets. In Figure 5f, certain cell populations, such as T and NK cells, CD4 T cells and CD8 T cells, and CD14 and CD16 Monocytes, are better separated in more distinct clusters than in Figure 5d.

Discussion

Sample barcoding, as implemented in the 10X Flex protocol, offers a solution to two significant issues in droplet-based scRNA-seq experiments: robust identification of sample-of-origin in multiplexed cell pools, and deconvolution of multiple cells from different samples in the same droplet. However, herein we demonstrate that unresolvable doublets, consisting of multiple cells from the same sample, remain a significant issue in Flex data sets, with scDblFinder classifying several times more cells as doublets than predicted by the manufacturer.

Despite the greater-than-expected number of doublets in Flex data, we find that in general the Flex protocol provides similar data quality to previous 3' scRNA-seq protocols, and existing tools and workflows, such as ambient RNA removal and cell annotation, work without issue. This includes doublet finders previously developed for 3' data sets.

We verify that the predictions of scDblFinder are accurate by aggregating the droplet counts across the sample barcodes, a novel approach that will prove useful in future benchmarking studies of doublet-finding algorithms. We then propose a mathematical model for predicting the number of same-sample doublets in any droplet-based single-cell protocol with sample barcodes, and find that the predictions of our model are much closer to the observed doublet rates in Flex experiments than the

estimates from the user documentation. Importantly, we demonstrate that doublet removal improves downstream analysis by eliminating spurious cell clusters consisting primarily of doublets.

Methods

Single-cell data generation

For the earlyAIR data set, samples were processed to single-cell suspensions, fixed within 1 hour of collection according to 10x Genomics Flex Protocol recommendations [5] and stored at -80C until enough samples for one batch of 16 had been collected. Samples were thawed, barcoded, pooled (16 samples per pool), and processed according to 10x Genomics recommendations without modifications. Fo r each capture, 128,000 cells were targeted, libraries were prepared using standard protocols, and sequencing was performed to achieve 10,000 reads/cell.

Single-cell data pre-processing strategy

The starting point for all data sets discussed in this paper is a counts matrix consisting of cells called by CellRanger. This data is pre-processed in the following order prior to running doublet-finding tools.

- Removal of unexpressed genes.
- Removal of ambient background contamination using decontX [17].
- Removal of cells with fewer than 250 reads.
- Removal of cells with high mitochondrial read percentages, using the `isOutlier` function in the `DropletUtils` package [7, 10].
- Doublet calling with scDblFinder [6]

Overview of doublet-finding tools

We test two doublet-finding tools on the droplet-level counts for the PBMC data set, scds and scDblFinder. scds [2] has two methods for doublet detection, the coexpression method and the binary classification. The co-expression method assigns a score to every pair of genes based on how often they are expressed in the same droplets. This score is aggregated across all pairs for each droplet to create a doublet score, which is higher for droplets that co-express gene pairs with a low co-expression score. The binary classification method generates artificial doublets by adding random pairs of columns in the counts matrix and creating an augmented data set consisting of the input data and the artificial doublets. A classifier for distinguishing the input data from the artificial doublets is trained on the combined data set using boosted decision trees, and then run on the input data alone to compute a doublet score for each droplet. scds can also compute a hybrid score which is the average of the scores from the co-expression and binary classification methods, and droplets are called either singlets or doublets based on a threshold determined from the expected number of doublets in the data set.

scDblFinder [6]works similarly to the binary classification method in scds but rather than using the counts matrix directly, artificial doublets are generated in a lower dimensional representation of the expression space and creating a k-nearest neighbours graph. scDblFinder runs for several iterations, removing the input droplets with the highest doublet scores and regenerating the artificial doublets at each iteration.

Data availability

Code for performing all analysis included within this paper, as well as a table of summarized data for each batch used to generate Figures 2 and 4 is available at `https://github.com/Oshlack/flex-doublets`. The 10X PBMC data set is available online at `https://www.10xgenomics.com/datasets/128k-human-pbmcs-stained-with-totalseqc-human-universal-cocktail`. The earlyAIR data set is currently under embargo but will be made available via CZ CELLxGENE.

Acknowledgments

Data generation for the earlyAIR data set was performed by WEHI's Cellular Genomics Projects Team, with sequencing performed by Stephen Wilcox and Sarah MacRaild and additional data generation by Casey J. A. Anttila and Ruvimbo Mishi. We thank Professor Andrew Melatos for suggesting using partitions as a way to describe the distribution of cells within droplets.

This project is supported by CZI Pediatric Networks for the Human Cell Atlas 2021-237883. AO is supported by NHMRC Investigator grant (GNT1196256). JM is supported by NHMRC Investigator grant (GNT1187748).

References

1. George E Andrews. *The theory of partitions*. Number 2. Cambridge university press, 1998.
2. Abha S Bais and Dennis Kostka. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics*, 36(4):1150–1158, February 2020.
3. Daniel V Brown, Casey J A Anttila, Ling Ling, Patrick Grave, Tracey M Baldwin, Ryan Munnings, Anthony J Farchione, Vanessa L Bryant, Amelia Dunstone, Christine Biben, Samir Taoudi, Tom S Weber, Shalin H Naik, Anthony Hadla, Holly E Barker, Cassandra J Vandenberg, Genevieve Dall, Clare L Scott, Zachery Moore, James R Whittle, Saskia Freytag, Sarah A Best, Anthony T Papenfuss, Sam W Z Olechnowicz, Sarah E MacRaild, Stephen Wilcox, Peter F Hickey, Daniela Amann-Zalcenstein, and Rory Bowden. A risk-reward examination of sample multiplexing reagents for single cell RNA-Seq. *Genomics*, 116(2):110793, March 2024.
4. Fabiola Curion, Xichen Wu, Lukas Heumos, Mylene Mariana Gonzales André, Lennard Halle, Matiss Ozols, Melissa Grant-Peters, Charlotte Rich-Griffin, Hing-Yuen Yeung, Calliope A Dendrou, Herbert B Schiller, and Fabian J Theis. Hadge: A comprehensive pipeline for donor deconvolution in single-cell studies. *Genome Biol.*, 25(1):109, April 2024.
5. 10X Genomics. *Chromium Fixed RNA Profiling Reagent Kits for Multiplexed Samples*. 10X Genomics, Pleasanton, CA, September 2023. Revision E, available at `https://www.10xgenomics.com/support/single-cell-gene-expression-flex/documentation/steps/library-prep/chromium-single-cell-gene-expression-flex-reagent-kits-for-multiplexed-samples`.
6. Pierre-Luc Germain, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D Robinson. Doublet identification in single-cell sequencing data using *scdblfinder*. *F1000Res.*, 10:979, September 2021.
7. Jonathan A Griffiths, Arianne C Richard, Karsten Bach, Aaron T L Lun, and John C Marioni. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.*, 9(1):2667, July 2018.
8. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

9. George Howitt, Yuzhou Feng, Lucas Tobar, Dane Vassiliadis, Peter Hickey, Mark A Dawson, Sarath Ranganathan, Shivanthan Shanthikumar, Melanie Neeland, Jovana Maksimovic, and Alicia Oshlack. Benchmarking single-cell hashtag oligo demultiplexing methods. *NAR Genomics and Bioinformatics*, 5(4):lqad086, 10 2023.
10. Aaron T L Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C Marioni. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, 20(1):63, March 2019.
11. Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, July 2019.
12. Ryan M Mulqueen, Dmitry Pokholok, Brendan L O'Connell, Casey A Thornton, Fan Zhang, Brian J O'Roak, Jason Link, Galip Gürkan Yardımcı, Rosalie C Sears, Frank J Steemers, and Andrew C Adey. High-content single-cell combinatorial indexing. *Nat. Biotechnol.*, 39(12):1574–1580, December 2021.
13. Drew Neavin, Anne Senabouth, Himanshi Arora, Jimmy Tsz Hang Lee, Aida Ripoll-Cladellas, sc-eQTLGen Consortium, Lude Franke, Shyam Prabhakar, Chun Jimmie Ye, Davis J McCarthy, Marta Melé, Martin Hemberg, and Joseph E Powell. Demuxafy: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biol.*, 25(1):94, April 2024.
14. Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z Yeung, William M Mauck, 3rd, Peter Smibert, and Rahul Satija. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.*, 19(1):224, December 2018.
15. Nan Miles Xi and Jingyi Jessica Li. Benchmarking computational Doublet-Detection methods for Single-Cell RNA sequencing data. *Cell Syst*, 12(2):176–194.e6, February 2021.
16. Ke-Xu Xiong, Han-Lin Zhou, Cong Lin, Jian-Hua Yin, Karsten Kristiansen, Huan-Ming Yang, and Gui-Bo Li. Chord: an ensemble machine learning algorithm to identify doublets in single-cell RNA sequencing data. *Commun Biol*, 5(1):510, May 2022.
17. Shiyi Yang, Sean E Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima, and Joshua D Campbell. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.*, 21(1):57, March 2020.
18. Ziyang Zhang, Madeline E Melzer, Keerthana M Arun, Hanxiao Sun, Carl-Johan Eriksson, Itai Fabian, Sagi Shaashua, Karun Kiani, Yaara Oren, and Yogesh Goyal. Synthetic DNA barcodes identify singlets in scRNA-seq datasets and evaluate doublet algorithms. *Cell Genom*, 4(7):100592, July 2024.