Journal Title Here, 2024, 1–6
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 R. 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-specific barcoding of each molecule in single-cell RNA-sequencing experiments allows deconvolution of RNA from different cells captured together within a droplet. 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 in 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 [13] or lipids [10] on the cell surface and can be sequenced along with the gene expression. Various methods have been developed to identify the sample of origin from the expression counts of the hashtags [8]. 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 [2]. Secondly, demultiplexing hashtag counts is a challenging computational problem, and the number of cells that can be confidently assigned to samples is dependent on the choice of demultiplexing software, especially for poor quality hashtagging [8].

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 [12]. 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 Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

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 multiplets increases steeply with the number of cells in individual droplets, which we refer to as the droplet occupancy. 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 [11]. 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 collected in the clinic from "healthy donors". Some donors provided multiple tissues, and the batches contain between 25,508 and 133,247 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 135,707 cells after pre-processing.

Measuring multiplet rates in Flex data

To identify doublets in our data, we use scDblFinder [6] which offers high performance and ease of use with relatively low computational requirements [12]. A brief discussion of how scDblFinder works and the settings used in this study are provided in the Methods section. Figure 1 shows the fraction of doublets called by scDblFinder on each 16-plex Flex run as a function of the number of cells. In general, the fraction of doublets is expected to increase linearly with the number of cells loaded [4], and the number of doublets increases quadratically with the number of cells. The documentation for the Flex protocol predicts a fraction of 0.8% per 1000 cells per sample, equivalent toor 0.05% per 1000 cells for 16-plex runs. While there is a large amount of scatter in the results from scDblFinder, the doublet fraction from scDblFinder is higher than the 10X value in all batches, with a fitted fraction of 0.12% per 1000 cells. Supplementary Figure 1 shows the doublet fraction of each sample in all batches as a function of the number of cells, and a modest positive correlation is observed in some batches.

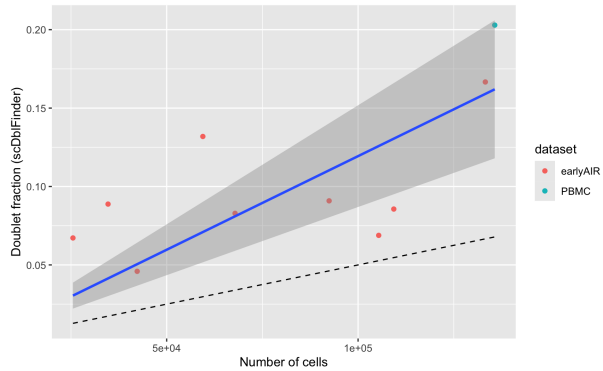

Fig. 1: Doublet fraction estimated from scDblFinder vs number of cells called per run for 10, 16-plex Flex runs. The blue line indicates a linear fit to the data. The black dashed line is the estimated rate from the Flex user guide (0.05% per 1000 cells).

Estimating the resolvable cell fraction

Why is the doublet rate from scDblFinder greater than the rate in the Flex user guide? In a 3' experiment, each column in the counts matrix corresponds to the RNA detected in a droplet. The doublet rate can be estimated by assuming a Poisson process where the rate parameter is the proportion of droplets containing cells. By contrast, when processing Flex data, the counts are split by sample barcode into 'pseudo-droplets' consisting of each droplet plus sample barcode combination, so the number of columns in the counts matrix is the number of physical droplets multiplied by the number of samples. In this section, we show that the number of same-sample doublets in a Flex experiment increases as a function of the mean droplet occupancy, i.e. the number of cells from *all* samples in *physical* droplets. If each sample is considered separately in pseudo-droplets, then the average occupancy measured is artificially lower, leading to an underestimation of the doublet rate. Figure 2 gives a heuristic example: on the left, 10 cells from four unique samples are captured within five droplets giving an average occupancy of 2. On the right, the same 10 cells are now spread across 9 (non-empty) pseudo-droplets, so that the measured occupancy 1.1 and so is roughly half the true occupancy.

We now show how the rate of same-sample doublets is a function of the droplet occupancy. 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,

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

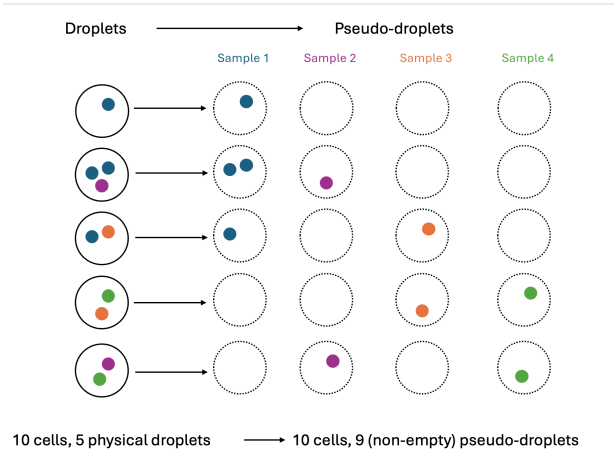

Fig. 2: A schematic depiction of how sample-barcoded cells in physical droplets (left) are split into sample-specific pseudo-droplets (right) in Flex data processing. Black circles with solid borders represent physical droplets, those with dashed borders represent pseudo-droplets. Small filled circles represent cells, with colours corresponding to unique samples. The pseudo-droplets have lower average occupancy than the physical droplets.

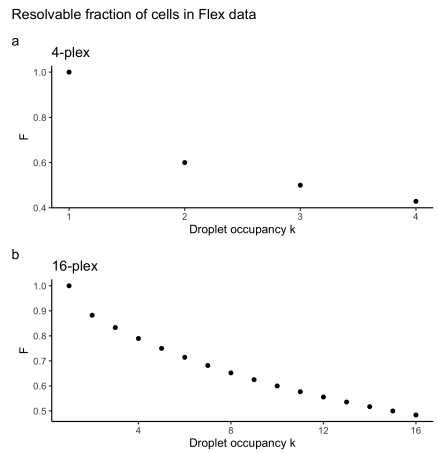

Fig. 3: Fraction of resolvable cells $F(n,k)$ vs occupancy k, where n is the number of multiplexed samples and k is the number of cells in each droplets. (a) $n = 4$ and (b) $n = 16$.

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

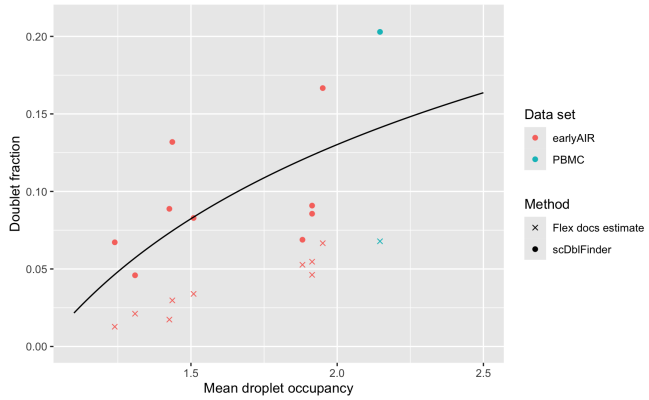

Fig. 4: Fraction of multiplets in Flex data vs mean droplet occupancy. Black curve is the calculated value from equation 12 for n = 16 samples. Coloured points indicate the doublet fraction determined by scDblFinder and observed mean droplet occupancy from the earlyAIR and PBMC data sets. Crosses indicate the predicted values from the Flex documentation based on the number of cells in each batch.

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

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

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], as well as the expected doublet fraction from the Flex documentation [5].

Figure 4 shows that the doublet fraction values from scDblFinder are consistently higher than those estimated from the Flex documentation. The measured doublet fractions from scDblFinder straddle the theoretical predictions from (12), and in all but one instance the scDblFinder values are closer to our predicted values than the values from the Flex documentation. However, some limitations to both our model and data remain. 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 may shift the data points in Figure 4 to the right, with the effect likely to be more pronounced in batches with higher average occupancy, (see Figure 3). Second, we assume that all samples have equal numbers of cells. In any experiment, some samples contain more cells than others. We expect that this affects the rate of same-sample multiplets, though exactly how is beyond the scope of this work. Third, homogenous tissues are likely to contain more homotypic doublets than tissues with more diverse cell populations. Doublet finders struggle to identify these doublets [12], leading to a lower than expected doublet fraction. Finally, some tissue types are more likely than others to contain 'biological' doublets, where two or more cells are physically linked and are therefore more likely to enter droplets at the same time, which may result in a higher than expected doublet rate [3].

Doublet-finding identifies spurious clusters

In order to demonstrate the effect of doublets on downstream analysis, we perform unsupervised clustering of the PBMC data set at a range of resolutions from 0.1 to 1.0 using the `FindClusters` function with the SLM algorithm in Seurat and count the number of clusters at each resolution. Figure 5a shows the results for three versions of the data: all cells (including doublets) that pass our initial QC filtering, singlets-only (after removing all of the doublets identified by scDblFinder), and downsampled cells. Removing the doublets always results in fewer clusters than keeping them in the analysis. However, the data from all cells contains 27,536 more cells than the singlets-only data, so the observation could be driven by the number of cells, rather than the presence or absence of doublets. To check this, the downsampled data is made by randomly removing an equivalent number of cells as the number of doublets from all cells, so that the resulting data set has the same number of cells as the singlets-only data with the same singlet/doublet ratio as the data from all cells. At every resolution, removing the doublets reduces the number of clusters, even for the versions of the data containing the same number of cells (Figure 5a), indicating that doublets introduce structure in the data. The randomly downsampled data is generated 10 times, the number of clusters displayed in Figure 5a are the mean values at each resolution.

We next look at the doublet composition of the clusters identified using all cells at resolution 0.4 (Figures 5b and 5c. We identify 9 clusters out of 30 with more than 50% doublets (the average doublet fraction for the whole data set is 20.2%). In the downsampled data sets, on average 7 out of 28 clusters are dominated by doublets, indicating that retaining doublets can result in annotation of dubious clusters. Figure 5d shows the position of all cells in the data set prior to doublet removal, as annotated by Azimuth using a PBMC reference, on a Uniform Manifold Approximation and Projection (UMAP) plot. Figure 5e shows only the cells in the 9 clusters identified as being composed of mostly doublets. Figures 5d and 5e highlight that the 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. Figure 5f shows the UMAP with Azimuth labels after removing the doublets. The average silhouette width of the data set with all cells is 0.158, which increases to 0.236 for the singlet-only data set, indicating that removal of doublets leads to better-differentiated clusters.

We next look at the doublet composition of the identified cluster at resolution 0.4, (Figures 5b and c). We identify 9 clusters out of 30 with more than 50% doublets (the average for the whole data set is 20.2%). In the downsampled data sets, on average 7 out of 28 clusters are dominated by doublets. 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

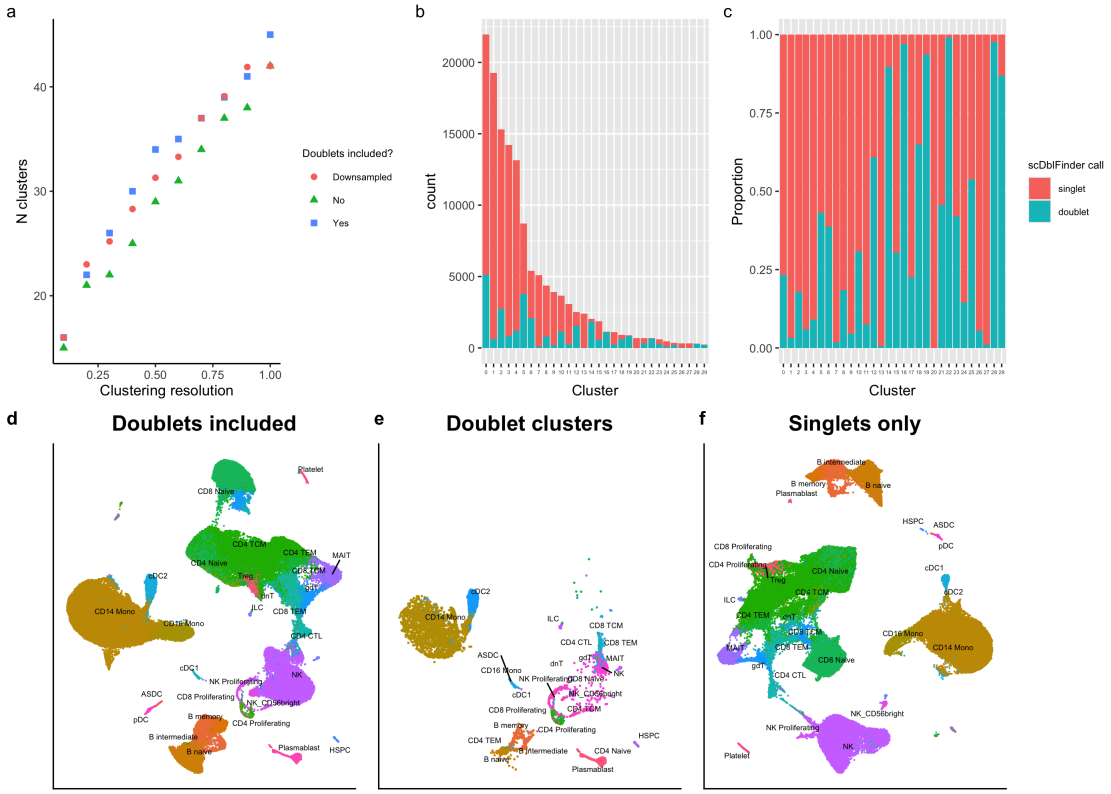

Fig. 5: a) Number of clusters as a function of cluster resolution for PBMC data set before (blue triangles) and after (red circles) removal of doublets identified by scDblFinder. Green triangles show the mean number of clusters from 10 downsampled data sets with doublets included but the same number of cells as the singlet-only data. 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 $\geq$ 75% doublets. f) UMAP with Azimuth labels computed after removing doublets.

in the 9 clusters 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.

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 of reads, 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 in the same droplet 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, our experience of analysing data from the Flex protocol results in 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 using doublet finders previously developed for 3′ data sets which we recommend as an important component of the analysis.

We 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 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. For 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 version 7.1.0. This data is pre-processed in the following order prior to running doublet-finding tools.

- Removal of all genes not detected in any droplet.
- Removal of ambient background contamination using decontX [14].
- Removal of cells with fewer than 250 UMIs.
- Removal of cells with high mitochondrial read percentages, using the `isOutlier` function in the `DropletUtils` package [7, 9].
- Doublet calling with scDblFinder [6]

Overview of scDblFinder

We use scDblFinder [6] throughout this paper to call doublets. scDblFinder works by generating artificial doublets by combining the expression of two cells in the input data, and creating an augmented data set consisting of the input data and the artificial doublets. It then performs a dimensional reduction and constructs a nearest neighbour network in the lower dimensional space, adding estimated and observed features of the data set. Gradient boosted trees are trained on this network to distinguish the input data from the artificial doublets, computing a doublet score for each cell in the input data. The classifier runs for several iterations, removing the input droplets with the highest doublet scores at each iteration, with the final threshold being determined by minimising misclassification of artificial doublets.

For the single-cell data sets, we use scDblFinder version 1.18.0with two exceptions, we use default values for all arguments. The first is `clusters = TRUE`. This option performs a fast clustering on the data, then generates artificial doublets by only combining cells from different clusters, rather than at random in its default. The second is `dbr.sd = 1`. By default, scDblFinder assumes a doublet rate of 1% per 1000 cells, based on the documentation for the 10X Flex 3′ protocol [4]. Setting `dbr.sd = 1` removes this assumption, and the doublet rate is determined solely by finding an optimal doublet score threshold to minimise misclassification of artificial doublets.

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 1 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. 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.
3. Jinguo Chen, Foo Cheung, Rongye Shi, Huizhi Zhou, Wenrui Lu, and CHI Consortium. PBMC fixation and processing for chromium single-cell RNA sequencing. *J. Transl. Med.*, 16(1):198, July 2018.
4. 10X Genomics. *Chromium Next GEM Single Cell 3' Reagent Kits v3.1 User Guide*. 10X Genomics, Pleasanton, CA, November 2019. Revision D, available at `https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/library-prep/chromium-single-cell-3-reagent-kits-user-guide-v-3-1-chemistry`.
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. 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.
9. 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.
10. 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.
11. 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.
12. 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.

13. 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.
14. 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.