On the correct detection of empty droplets in droplet-based single-cell RNA sequencing protocols

Aaron T. L. Lun[1,*], Samantha Riesenfeld[2,*], Tallulah Andrews[3,*], The Phuong Dao[4,*], Tomas Gomes[3,*] and others

1 Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom
2 Something... Broad?
3 Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom
4 Something... Columbia?

*** These authors contributed equally to this work.**

Introduction

Recent advances in droplet-based protocols have revolutionized the field of single-cell transcriptomics by allowing tens of thousands of cells to be profiled in a single assay [1–3]. In these technologies, individual cells are captured into aqueous droplets in a water-in-oil emulsion. Each droplet also contains a co-captured bead with primers for reverse transcription, where all primers on a single bead contain a cell barcode that is (effectively) unique to that bead. The droplets serve as isolated reaction chambers in which cell lysis and reverse transcription are performed to obtain barcoded cDNA. This is followed by breaking of the emulsion, amplification of the cDNA and construction of a sequencing library. After sequencing, debarcoding is performed based on the cell barcode observed in each read sequence. This yields an expression profile for each cell, typically in the form of unique molecular identifier (UMI) counts [4] for all annotated genes. The use of droplets increases throughput by at least an order of magnitude compared to protocols based on plates [5] or conventional microfluidics [6], which is appealing for large-scale projects such as the Human Cell Atlas [7].

That said, the complexity of the sequencing data from droplet-based technologies poses a number of interesting challenges for low-level data processing. One such challenge is the identification and removal of cell barcodes corresponding to empty droplets. An empty droplet does not contain a cell but will still contain "ambient" RNA [1], i.e., cell-free transcripts in the solution in which the cells are suspended. Ambient RNA may be actively secreted by cells or released upon cell lysis, the latter of which is particularly likely given the stresses of dissociation. The presence of ambient RNA means that many empty droplets will contain material for reverse transcription and library preparation, resulting in non-zero total UMI counts for the corresponding barcodes. However, the resulting expression profiles do not originate from any single cell and need to be removed prior to further analysis to avoid misleading conclusions.

Existing methods for removing empty droplets assume that droplets containing genuine cells should have more RNA, resulting in larger total UMI counts for the corresponding barcodes. Zheng *et al.* [3] remove all barcodes with total counts below 10% of the 99^{th} percentile of the Y largest total counts (where Y is defined as the expected number of cells to be captured on the Chromium device). Macosko *et al.* [1]

set the threshold at the inflection point in the cumulative fraction of reads with respect to increasing total count. While simple, the use of a one-dimensional filter on the total UMI count is suboptimal as it may discard small cells with low RNA content. Droplets containing small cells may not be easily distinguishable from large empty droplets in terms of the total number of transcripts. This problem is exacerbated by variable capture and amplification efficiency across droplets, which further mixes the distributions of total counts between empty and non-empty droplets. A simple threshold on the total count forces the researcher into a difficult choice between the loss of small cells or an increase to the number of artifactual "cells" composed of ambient RNA.

In this report, we propose a new method for detecting empty droplets in droplet-based single-cell RNA sequencing (scRNA-seq) data. We construct a profile of the ambient pool of RNA, and test each barcode for deviations from this profile using a Poisson-based model for the count distribution. Barcodes with significant deviations are considered to be genuine cells, thus allowing recovery of cells with low total RNA content and small total UMI counts. We combine our approach with an inflection point filter to ensure that barcodes with large total counts are always retained. Using a variety of simulations, we demonstrate that our method outperforms any simple threshold on the total UMI count. We also apply our method to several real data sets where we are able to recover more cells from both existing and new cell types.

Description of the method

Testing for deviations from the ambient profile

To construct the profile for the ambient RNA pool, we consider a threshold T on the total UMI count. The set $\mathcal{D}$ of all barcodes with total counts less than T are considered to represent empty droplets. The exact choice of T does not matter, as long as (i) it is small enough so that droplets with genuine cells do not have total counts below T, and (ii) there are sufficient counts to obtain a precise estimate of the ambient profile. A simple approach is to define empty droplets from all but the top 100,000 barcodes with the largest total counts, under the assumption that fewer than 100,000 cells are captured in any run. We stress that T is not the same as the threshold used in existing methods, as we make no statement on barcodes with total counts greater than T at this point.

The ambient profile is constructed by summing counts for each gene across $\mathcal{D}$. Let y_{gb} be the count for gene g in barcode b. We define the ambient count as

$$A_g = \sum_{b \in \mathcal{D}} y_{gb} \ ,$$

yielding a count vector $\mathbf{A} = (A_1, \ldots, A_N)$ for all N genes. (We assume that any gene with counts of zero for all barcodes has already been filtered out, as this provides no information for distiguishing between barcodes.) We apply the Good-Turing algorithm to $\mathbf{A}$ to obtain the posterior expectation $\tilde{p}_g$ of the proportion of counts assigned to g [8], using the `goodTuringProportions` function in the `edgeR` package [9]. This ensures that genes with zero counts in the ambient pool have non-zero proportions, avoiding the possibility of obtaining likelihoods of zero in downstream calculations.

Each barcode with a total count above T is then fitted to this ambient profile. Consider a barcode b with a total count t_b, corresponding to an empty droplet. For this barcode, we assume that the count for each gene follows a Poisson distribution with mean $\lambda_{gb} = \tilde{p}_g t_b$. This is based on random sampling of free-floating transcripts in solution into the empty droplets. We also assume that the sampling procedure is

independent between genes. The deviance of the fit for this barcode is written as

$$D_b = 2\left[\sum_g y_{gb}\log\left(\frac{y_{gb}}{\lambda_{gb}}\right) - y_{gb} + \lambda_{gb}\right],$$

where the sum is taken across all genes with at least one non-zero count.

Based on generalized linear model theory, D_b should approximately follow a χ^2_N distribution under the null hypothesis, i.e., if the barcode truly originates from an empty droplet. One could then reject the null hypothesis by computing the upper tail probability of the χ^2_N distribution at D_b. However, this approach is not reliable in practice, possibly due to the failure of the saddlepoint approximation at low counts. We instead use a Monte Carlo approach to determine the distribution of D_b under the null:

1. Let the largest value of t_b be t_M. Define a tolerance τ, and split the range $[\log_2(T) - \tau, \log_2(t_M) + \tau]$ into S equidistant points. S should be chosen such that the interval $[\log_2(t_b) - \tau, \log_2(t_b) + \tau]$ for any b contains $R = 10^5$ points.

2. Let point s have a total count of t^*_s. For each gene g, randomly sample a count from a Poisson distribution with mean $\tilde{p}_g t^*_s$. Compute the deviance D^*_s from the resulting count vector, using the same expression as described for D_b.

3. Fit a trend to D^*_s against $\log_2(t^*_s)$ for all points. We use a loess smoother with degree 1 and span 0.2, though any smoothing algorithm can be used. This yields a function $f(.)$, which returns the expected deviance at a given log-total. We also obtain the residuals from the trend, which we denote as r^*_s.

4. For each barcode b, compute the deviation from the trend as $r_b = D_b/f(\log_2(t_b))$. We identify all points with $\log_2(t^*_s)$ values in the interval $[\log_2(t_b) - \tau, \log_2(t_b) + \tau]$. Denote R_b as the number of points in this interval where $r^*_s \geq r_b$. We use the method of Phipson and Smyth [10] to compute a p-value for b as

$$P_b = \frac{R_b + 1}{R + 1}$$

This approach allows us to obtain permutation p-values for each barcode in a computationally efficient manner. R and τ determine the trade-off between speed and accuracy and can be set to arbitrarily large and small values, respectively, if computation time is no issue. In particular, larger values of R improve the precision with which P_b is calculated, while smaller values of τ improve accuracy in the presence of heteroskedasticity in the distribution of D^*_s with respect to increasing t^*_s.

Detecting the knee point in the log-totals

The procedure described above will identify barcodes that have count profiles that are significantly different from the ambient pool of RNA. This will be the case for most cell-containing droplets, as the ambient pool is formed from many (lysed) cells and is unlikely to be representative of any single cell. However, it is possible for some cell-containing droplets to have ambient-like expression profiles. This can occur if the cell population is highly homogeneous or if one cell subpopulation contributes disproportionately to the ambient pool, e.g., if it is more prone to lysis. Sequencing errors in the cell barcodes may also bias the estimates of the ambient proportions, by misassigning counts from cell-containing droplets to barcodes with low UMI totals. This may result in spurious similarities between cells and the estimated ambient profile.

If we apply our procedure directly, barcodes corresponding to ambient-like cell-containing droplets will be incorrectly filtered out. To avoid this, we combine our

procedure with a conventional threshold on the total UMI count. We rank all barcodes in order of decreasing t_b, and consider the function $f(.)$ of $\log(t_b)$ with respect to increasing log-rank. The first "knee" point in this function corresponds to a transition between a distinct subset of barcodes with large totals and the majority of barcodes with smaller totals. This is defined as the log-rank that maximizes the curvature

$$\frac{|f''|}{(1+f'^2)^{1.5}},$$

and represents the point at which $f(.)$ begins to drop rapidly, marking the start of the transition between large and small totals. We use the knee point rather than the inflection point as the former is more conservative, thus avoiding empty droplets.

Our assumption is that all barcodes with large totals must represent cell-containing droplets, regardless of whether its count profile resembles the ambient pool. This is based on the expectation that fluctuations in the size or capture efficiency of empty droplets should be smooth and not result in a distinct distribution of large totals. We define the upper threshold U as the t_b at the knee point and retain all barcodes with $t_b \geq U$, regardless of their P_b. This ensures recovery of barcodes corresponding to large cell-containing droplets. We stress that this approach is different from existing methods due to the use of our testing procedure. Barcodes with t_b below the knee point can still be retained if the count profile is significantly different from the ambient pool. This is not be possible with existing methods that would simply discard these barcodes.

Correcting for multiple testing across barcodes

We correct for multiple testing by controlling the false discovery rate (FDR) using the Benjamini-Hochberg method [11]. Putative cells are defined as those that have significantly poor fits to the ambient model at a specified FDR threshold. We set the FDR threshold to 1% by default, meaning that the expected proportion of empty droplets in the set of retained barcodes is no greater than 1%. Note that we only perform the correction on the p-values for barcodes that have t_b greater than T. This reduces the severity of the correction given that barcodes with lower t_b will always be discarded. Similarly, all barcodes with $t_b \geq U$ have their p-values set to zero during correction, as these barcodes are considered to be known true positives.

Results

Performance on simulated data

We named our method "EmptyDrops" and proceeded to test it on simulated data. Simulations were performed by mixing two real droplet-based scRNA-seq data sets involving cells with different RNA content (see Methods). This resulted in a single data set containing two groups of cells with low and high RNA content, and an ambient pool of RNA composed of a mixture of transcripts from both data sets. EmptyDrops was consistently able to detect cells from both groups with a small FDR (Table 1). In contrast, applying a threshold on the total UMI count was less effective at recovering cells from the smaller group. This is because barcodes corresponding to small cells with little RNA have similar total UMI counts as barcodes corresponding to large empty droplets with high levels of ambient RNA. The total UMI count cannot distinguish between these two possibilities, resulting in either reduced recall for the low-RNA group or a high false positive rate. This is demonstrated by the poor performance of the strategy used by 10X Genomics in their CellRanger software (Figures 1, 2).

Table 1. Performance of EmptyDrops on each simulation scenario with varying numbers of real cell barcodes from Group 2 (low RNA content). The recall for each group represents the proportion of real cells that were detected by EmptyDrops. The false discovery rate represents the proportion of detected barcodes that were empty droplets.

Scenario	**Group 1**		**Group 2**		**FDR (%)**
	Number	*Recall (%)*	*Number*	*Recall (%)*	
1	3285	99.85	205	96.10	0.00
2	3285	99.85	432	93.52	0.00
3	3285	99.88	815	93.74	0.01
4	3285	99.88	1731	93.41	0.00
5	3285	99.88	203	99.01	0.01
6	3285	99.85	391	99.74	0.00
7	3285	99.85	855	99.77	0.00
8	3285	99.88	1708	99.41	0.00
9	3285	99.85	180	100.00	0.00
10	3285	99.88	430	100.00	0.00
11	3285	99.85	810	100.00	0.00
12	3285	99.88	1644	99.94	0.00
13	3285	99.85	211	100.00	0.00
14	3285	99.85	445	100.00	0.00
15	3285	99.88	886	100.00	0.00
16	3285	99.85	1702	100.00	0.00

Performance on real data

We applied EmptyDrops to one channel of the 10X peripheral blood mononuclear cell (PBMC) data set [3]. EmptyDrops was able to identify barcodes with low UMI totals, which would have been discarded by the CellRanger threshold (Figure 3). This reflects the differences in performance observed in our simulations. Droplets containing small cells are indistinguishable from large empty droplets in terms of their total UMI counts, instead requiring a comparison of the expression profile to the ambient pool. Conversely, CellRanger detects a number of cells with large UMI totals that are not detected by EmptyDrops. This is attributable to the conservativeness of the knee-point threshold in EmptyDrops, which ensures that empty droplets are not inadvertently retained.

To explore the differences between methods in more detail, we generated a *t*-stochastic neighbour embedding (t-SNE) plot [12] of all barcodes that were detected by either method. We observed that the CellRanger-only barcodes clustered with barcodes that were detected by both methods (Figure 4). This suggests that the conservativeness of EmptyDrops can be largely ignored, as it only results in the loss of some cells from a cluster that would have been detected anyway. In contrast, the EmptyDrops-only barcodes formed three of their own clusters, such that the use of CellRanger would have resulted in the loss of entire groupsof cells. These clusters were associated with 101, 28 and 2339 significant marker genes, respectively (from left to right). The last group of cells we identified as platelets based on the expression of PF4 and PPBP (Figure 4). This is not surprising as the total RNA content of a cell is often associated with its type/state. The ability of EmptyDrops to retain small cells means that it can capture biology that would have been lost with CellRanger.

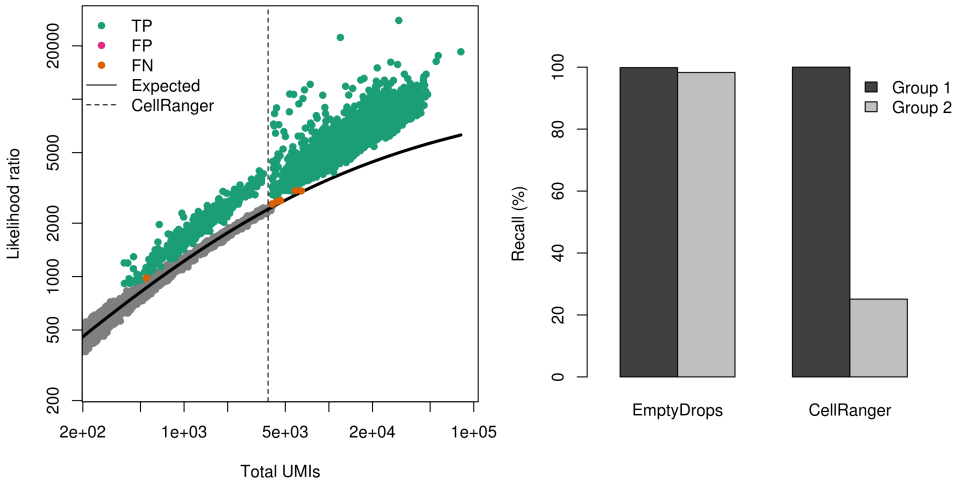

Figure 1. Behaviour of EmptyDrops and CellRanger in simulation scenario 6. Left: deviance statistics for each barcode in the simulated data, plotted against the total UMI count. Each barcode is represented by a point that is coloured according to whether it is a true positives, false positives or false negatives according to our method. The expected deviance under the null is also shown, along with the threshold defined by CellRanger. Right: recall of EmptyDrops and CellRanger for the two groups of cells in the simulated data. Group 1 contains the cells with more RNA while group 2 contains smaller cells.

Discussion

Droplet-based technologies are becoming increasingly popular for high-throughput single-cell transcriptomics. However, little work has been performed to develop computational methods for distinguishing genuine cells from empty droplets. Here, we describe EmptyDrops, a method to detect cell-containing barcodes based on significant deviation of the expression profiles from the pool of ambient RNA. We use simulated data to demonstrate that EmptyDrops outperforms the strategy that is currently implemented in the CellRanger software suite. Furthermore, EmptyDrops can recover biology in real 10X data that is lost using the CellRanger strategy. Our results indicate that EmptyDrops is effective for cell detection in droplet-based scRNA-seq data.

A key assumption of our approach is that barcodes with very low UMI totals represent empty droplets. This allows us to use these barcodes to obtain an estimate of the ambient pool. However, this assumption may not be appropriate if the data set contains a subset of cells with very low RNA content. In such cases, the estimate of the ambient expression profile will be biased, though this bias is likely to be small as relatively few transcripts will be contributed by cells with low RNA content. Another potential source of bias may arise from sequencing errors in the cell barcode, such that transcripts from a cell-containing droplet are misassigned to an otherwise empty droplet. This is mitigated to some extent by the use of designed cell barcodes in the GemCode protocol, which allow for error correction in the barcode [3]. However, it may be more of a problem in protocols where error correction of the barcodes is not possible [1].

An interesting question is whether the output of EmptyDrops is compatible with existing scRNA-seq analysis workflows. In particular, many workflows recommend the removal of cells with low total numbers of expressed genes during quality control [13, 14]. This would potentially result in the loss of cells that are recovered by EmptyDrops, defeating the purpose of using EmptyDrops in the first place. Moreover, it remains to

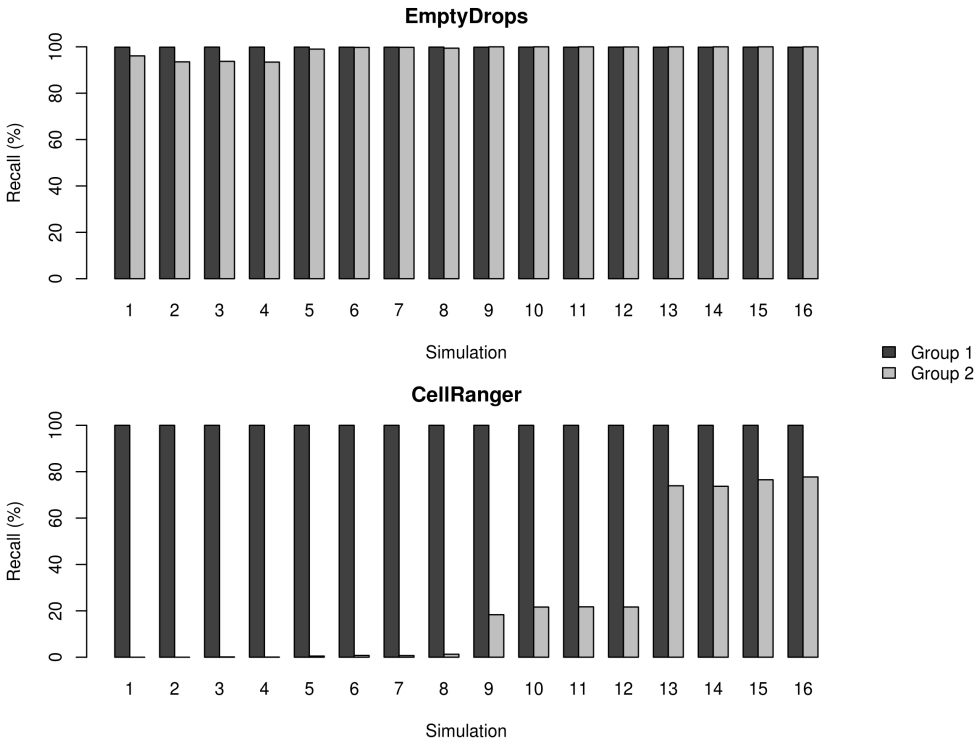

Figure 2. Comparison of recall between EmptyDrops (top) and CellRanger (bottom) in all simulation scenarios. Groups 1 and 2 contain larger and smaller cells, respectively.

be seen whether downstream analyses can effectively model the very high frequency of zero counts in these small cells. With improved recovery of small cells from droplet data, modifications may be required to existing scRNA-seq analysis pipelines.

Methods

Design of the simulations

We obtained raw barcode counts for two 10X data sets with XXX (high RNA content) and YYY cells (low RNA content). In each data set, we set a threshold on the total UMI count using the CellRanger approach. All barcodes with total counts above this threshold were considered to represent real cells, while all barcodes below this threshold were considered to represent empty droplets. Simulated data were generated by randomly sampling a number of real cell barcodes from the XXX and YYY data sets, which were treated as known real cells. To generate known empty droplets, we pooled the UMI counts for all empty droplet barcodes from both data sets. For each empty droplet barcode in the original data sets, we sampled with replacement from the UMI pool to obtain a new count profile with the same total UMI count. These were added to the simulated data to represent known empty droplets. Different simulation scenarios were constructed by varying the number of real cell barcodes that were sampled from each group.

We applied our EmptyDrops method to the simulated data to evaluate its performance. The recall was defined as the proportion of known real cells from each group that were successfully detected by our method. The observed false discovery rate

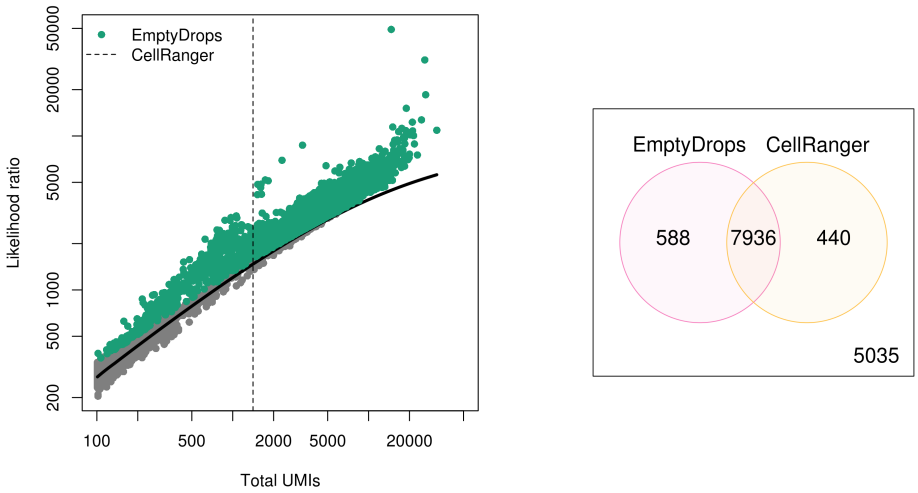

Figure 3. Behaviour of EmptyDrops and CellRanger on the 10X PBMC data set. Left: deviance statistics for all barcodes in real data, where each barcode is coloured depending on whether it was detected by EmptyDrops. The threshold defined by CellRanger is also shown. Right: numbers of barcodes detected by either or both methods.

was defined as the proportion of detected barcodes that were known empty droplets. We repeated this using the CellRanger approach, implemented as previously described [3].

Analysis of the 10X PBMC data set

We obtained one channel of the 10X PBMC data set from ???, and defined cells from the raw barcode counts using the EmptyDrops and CellRanger methods. For the CellRanger method, the expected number of cells was estimated as the number of cells identified at a FDR of 1% using EmptyDrops (8524 cells). This ensured a fair comparison between the two methods, though CellRanger was mostly robust to this parameter with 8318, 8381, and 8387 cells identified when the expected number of cells used was 5000, 8524, and 10000, respectively. For the set of barcodes detected by either method, we generated a t-SNE plot from the library-size normalized expression profiles using the Rtsne package (default parameters, perplexity of 30), using the top 500 highly variable genes. Library size normalization was performed by scaling the counts for each cell such that its total count was equal to the median total counts across all cells. Three groups enriched in cells identified only using EmptyDrops were segmented by eye, and marker genes were identified with a one-vs-all Wilcox-rank-sum test using a 5% FDR threshold.

References

1. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015.

2. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell

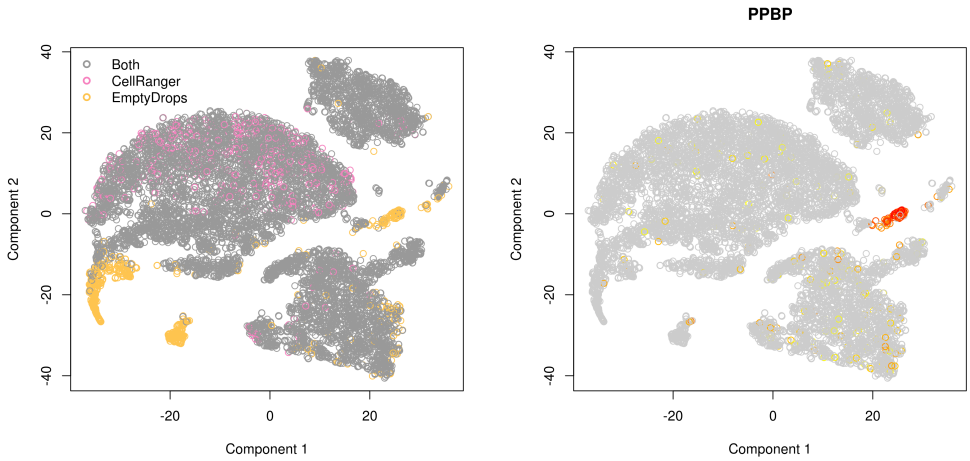

Figure 4. t-SNE plot of barcodes detected by either EmptyDrops or CellRanger. Each barcode is represented by a point, and is coloured based on whether it was detected by both methods (grey), EmptyDrops only (yellow) or CellRanger only (pink).

transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

3. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, 8:14049, Jan 2017.

4. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lonnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, Feb 2014.

5. S. Picelli, A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1098, Nov 2013.

6. A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, and J. A. West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, 32(10):1053–1058, Oct 2014.

7. A. Regev, S. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Gottgens, N. Hacohen, M. Haniffa, M. Hemberg, S. K. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, J. Lundeberg, P. Majumder, J. Marioni,

M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Philipakis, C. P. Ponting, S. R. Quake, W. Reik, O. Rozenblatt-Rosen, J. R. Sanes, R. Satija, T. Shumacher, A. K. Shalek, E. Shapiro, P. Sharma, J. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, A. van Oudenaarden, A. Wagner, F. M. Watt, J. S. Weissman, B. Wold, R. J. Xavier, and N. Yosef. The human cell atlas. *bioRxiv*, 2017.

8. William A Gale and Geoffrey Sampson. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

9. M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

10. B. Phipson and G. K. Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*, 9:Article39, 2010.

11. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, pages 289–300, 1995.

12. L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *J Mach Learn Res*, 9(2579-2605):85, 2008.

13. A. T. Lun, D. J. McCarthy, and J. C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*, 5:2122, 2016.

14. D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, Apr 2017.