Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data

Supplementary Materials

by

Aaron T. L. Lun[1,*,†], Samantha Riesenfeld[2,*], Tallulah Andrews[3,*], The Phuong Dao[4,*], Tomas Gomes[3,*], participants in the 1st Human Cell Atlas Jamboree[‡], John C. Marioni[1,3,5,#]

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom

[2]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

[4]Program for Computational and Systems Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY

[5]EMBL European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

* These authors contributed equally to this work.
† Email: aaron.lun@cruk.cam.ac.uk
‡ The full list of participants is provided in Supplementary Table 1.
# Email: marioni@ebi.ac.uk

March 28, 2018

1 Motivating the choice of the total count threshold T

The threshold T should be chosen so that cell-containing droplets are not used to estimate the ambient profile. Otherwise, our estimate will be distorted and we will have less power to discriminate between cells and empty droplets. To check if this occurs in real data, we calculated a p-value against the ambient null hypothesis for each barcode $b \in \mathcal{G}$, i.e., with $t_b \leq T$ where $T = 100$ by default. Ideally, we should observe a uniform distribution of p-values under the assumption that all barcodes in $\mathcal{G}$ are genuinely empty. However, if cell-containing droplets are present in $\mathcal{G}$, we should observe an enrichment of low p-values. These low p-values either correspond to the cell-containing droplets themselves, or to genuinely empty droplets that no longer fit to our distorted estimate of the ambient profile.

In most datasets, we observed a minor enrichment of low p-values in an otherwise uniform distribution (Supplementary Figure 1). This demonstrates that our testing procedure mostly holds its size and suggests that distortions of the ambient profile due to cell-containing droplets are not a major issue when using $T = 100$. The exception is the `neurons_900` dataset where we see a clear enrichment of very low and very large p-values. Note that this effect is not consistent with distortion of the ambient profile, which should only result in a skew towards low p-values. Rather, we hypothesize that it is driven by violations of the independent sampling assumption (e.g., if transcript molecules are complexed and sampled together into droplets). Positive correlations between molecules would increase the probability of obtaining droplets that are very similar or very different to the ambient profile.

2 Efficient calculation of the Monte Carlo p-value

The naive approach to computing the Monte Carlo p-value for a barcode would be to directly sample R count vectors from a multinomial distribution, compute the likelihood L'_{bi} for each count vector, and count the number of vectors with $L'_{bi} \leq L_b$. For N genes, we would need RN sampling operations to obtain the count vectors and the same number of multiplication operations to compute the likelihoods. For a dataset containing B barcodes, this would amount to a total of RNB operations.

When dealing with multiple barcodes, we can improve efficiency by using pre-existing results for barcodes with lower t_b. Within a simulation iteration i, we sample a count vector and calculate the likelihood $L'_{b_1 i}$ for the barcode b_1 with the lowest total count. For barcode b_2 with the total $t_{b_2} = t_{b_1} + 1$, we update the count vector by sampling a new gene k from the multinomial distribution with a size of 1 and probabilities equal to our ambient proportions. Similarly, we update the likelihood

$$L'_{b_2 i} = L'_{b_1 i} \frac{t_{b_2} \tilde{p}_k}{y'_{ki}} \ ,$$

where y'_{ki} is the frequency of k in the updated count vector. (Iterative application of these updates can be used in cases where $t_{b_2} > t_{b_1} + 1$.) The updated $L'_{b_2 i}$ can be directly compared to L_{b_2}, and repeating this procedure for R iterations allows us to compute the p-value for each b. This approach only requires $R \max\{t_b\}$ sampling and multiplication operations, and for droplet-based data, the maximum t_b across all barcodes in a dataset is much lower than NB. The former typically has values of 5000-10000, while the number of detected genes N is around 3000-5000 and B for all $t_b > T$ is often over 1000. This results in a considerable reduction in computational work compared to the naive approach.

Our algorithm can be viewed as an extension of a more simple speed-up where the set of likelihoods from all iterations are re-used for all barcodes with the same t_b. As a result, though, the Monte Carlo p-values are not strictly independent across barcodes. Instead, they are positively correlated as the same sampling choices are effectively shared between barcodes. This poses a theoretical problem for multiple testing correction with the Benjamini-Hochberg (BH) method, which relies on independent p-values. In practice, this is not a major issue as the BH method is robust to dependencies [1, 2]. Moreover, increasing R reduces the imprecision of the p-values caused by random sampling. This means that the effect of correlated estimation errors between different p-values will become negligible at large R. We use $R = 10000$ by default, though this can be increased if the lower bound on the p-values [3] is greater than the threshold for significance after the Benjamini-Hochberg correction.

References

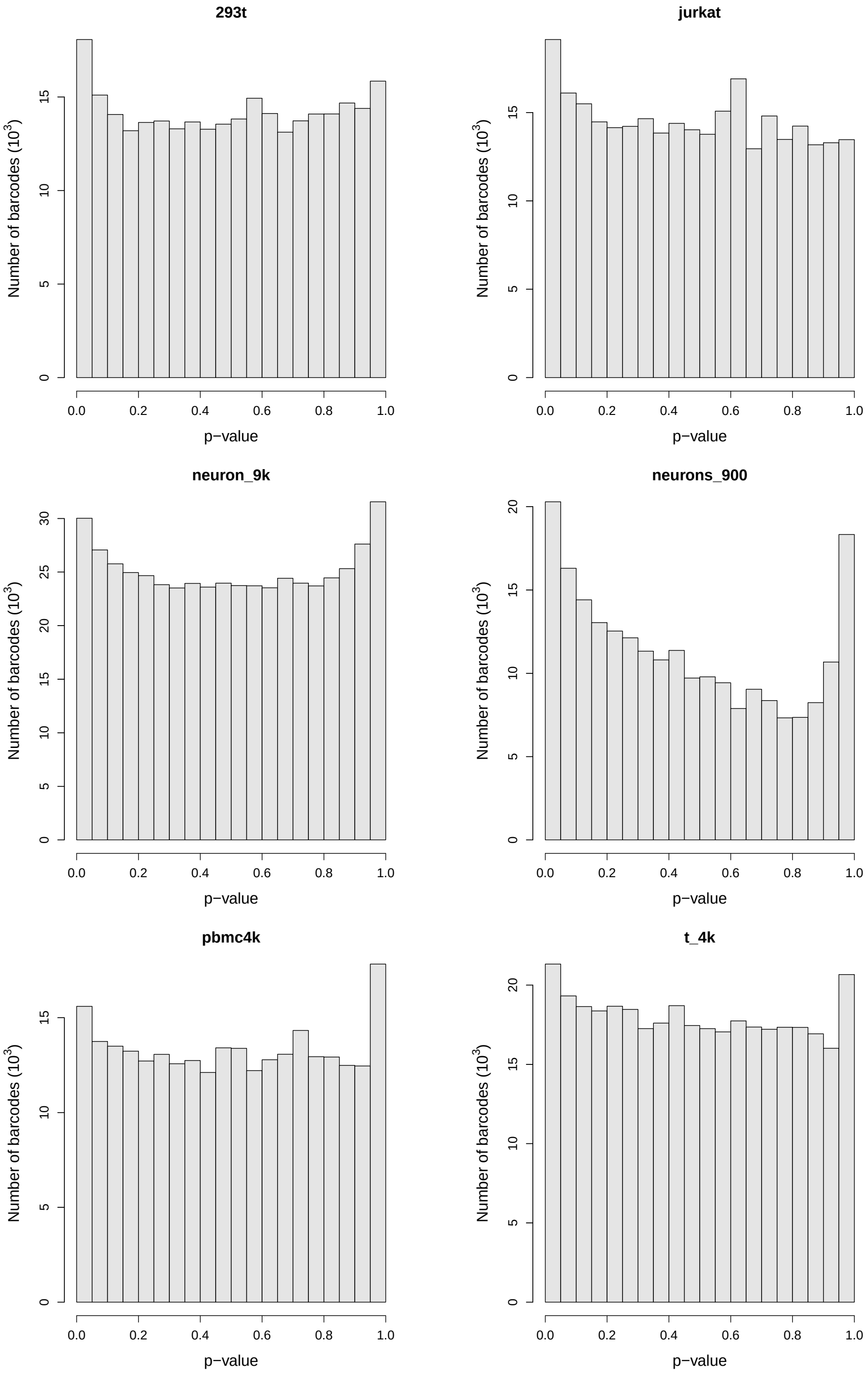
[1] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, Feb 2003.

[2] K. I. Kim and M. A. van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9:114, Feb 2008.

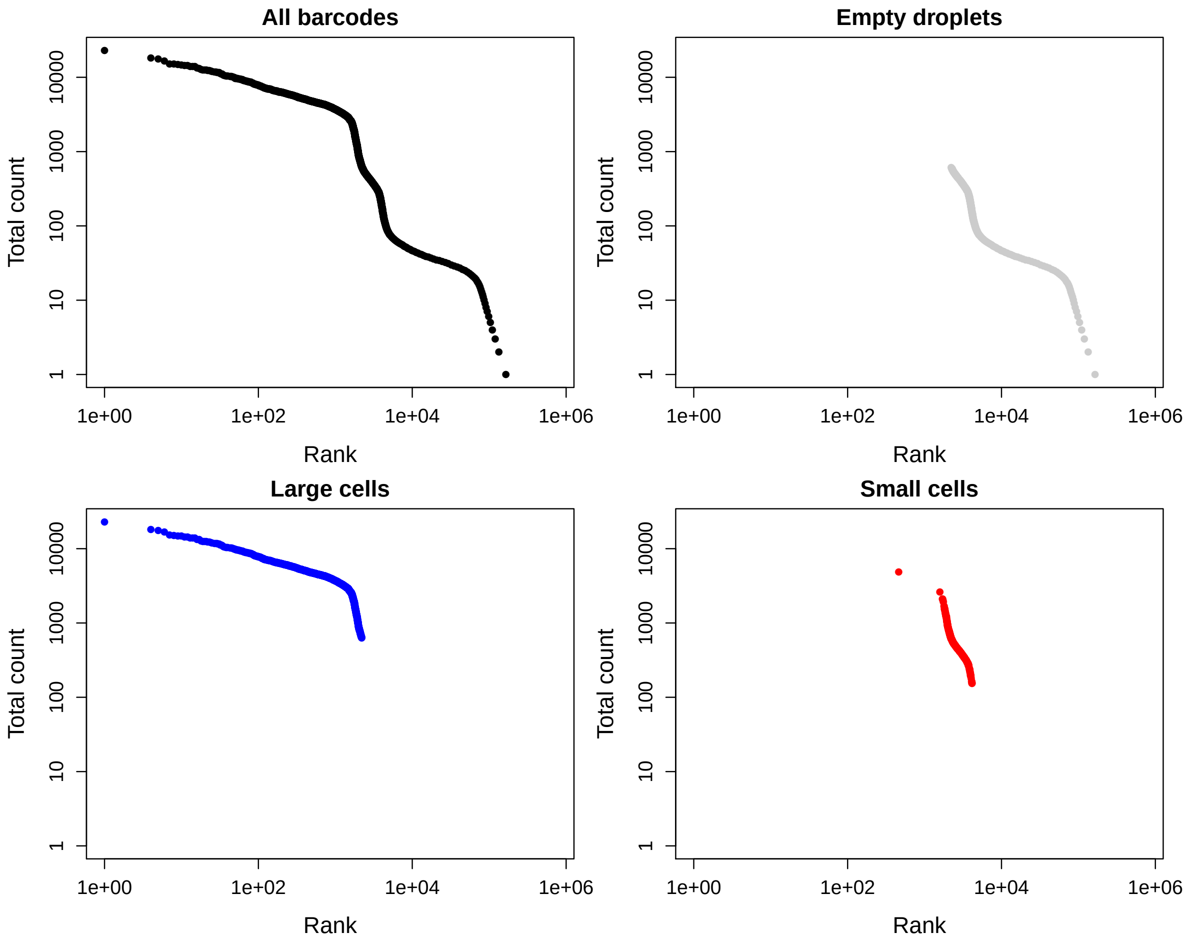
[3] B. Phipson and G. K. Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*, 9:Article39, 2010.

Supplementary Table 1: Summary of the datasets used to assess the various cell detection methods. All datasets were obtained from the 10X Genomics website. The organism, cell type, expected number of cells and the version of the CellRanger software used is shown for each dataset.

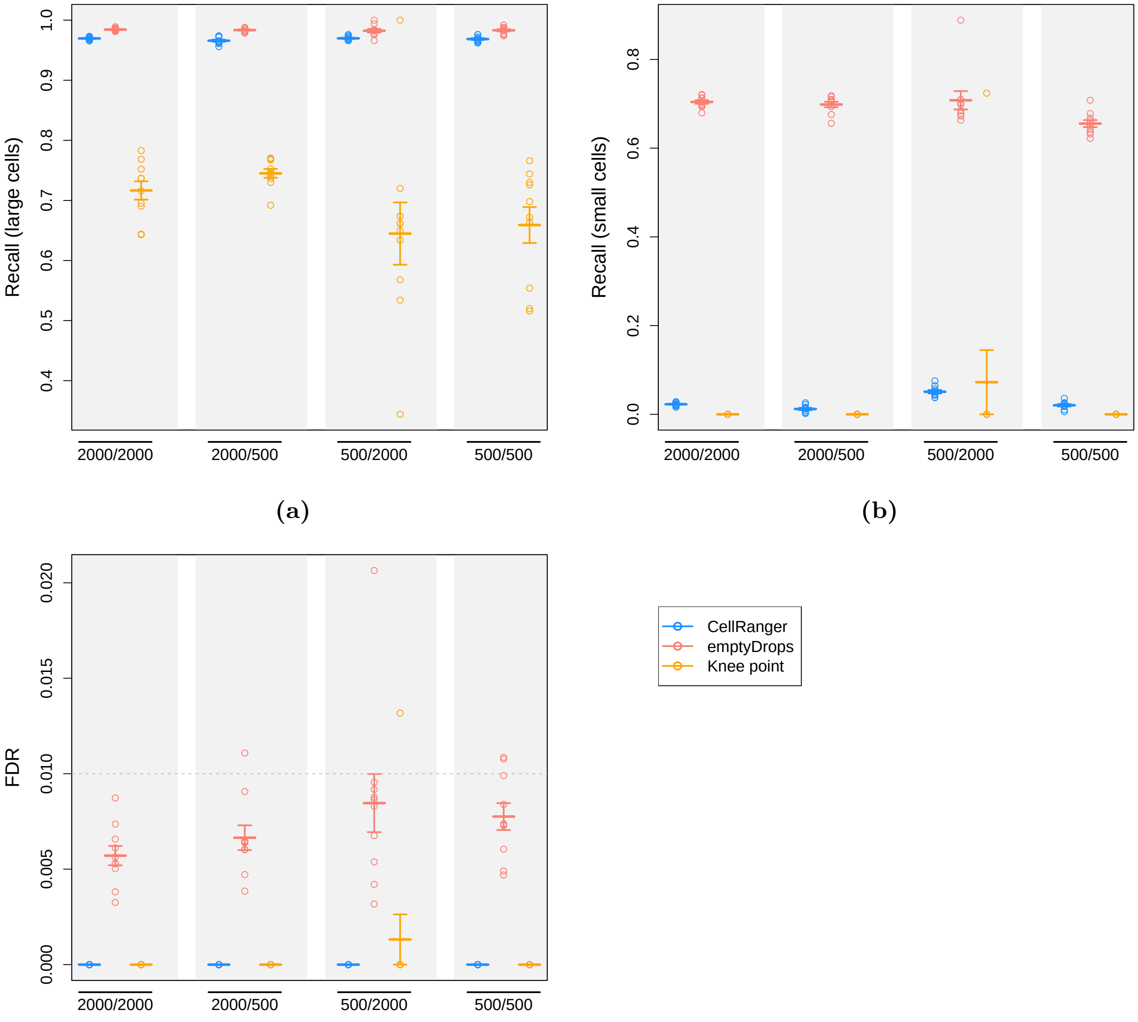
Name	**Organism**	**Cell type**	**Number**	**Version**
`pbmc4k`	Human	PBMCs	4000	2.1.0
`293t`	Human	293T cell line	2800	1.1.0
`jurkat`	Human	Jurkat cell line	3200	1.1.0
`neuron_9k`	Mouse	Brain cells	9000	2.1.0
`neurons_900`	Mouse	Brain cells	900	2.1.0
`t_4k`	Human	Pan T cells	4000	2.1.0



Supplementary Figure 1: Histograms of p-values for all barcodes with total UMI counts less than or equal to the threshold $T = 100$. Each plot corresponds to a dataset in Supplementary Table 1 where the p-value represents the deviation from the ambient profile.



Supplementary Figure 2: Total count against the rank for each barcode in a simulation based on the PBMC dataset with $G_1 = G_2 = 2000$. Plots are shown for all barcodes, barcodes corresponding to empty droplets, and barcodes corresponding to large or small cells. Ranks are calculated from the entire set of barcodes in all plots, for ease of comparison between plots. All axes are on a log-scale.

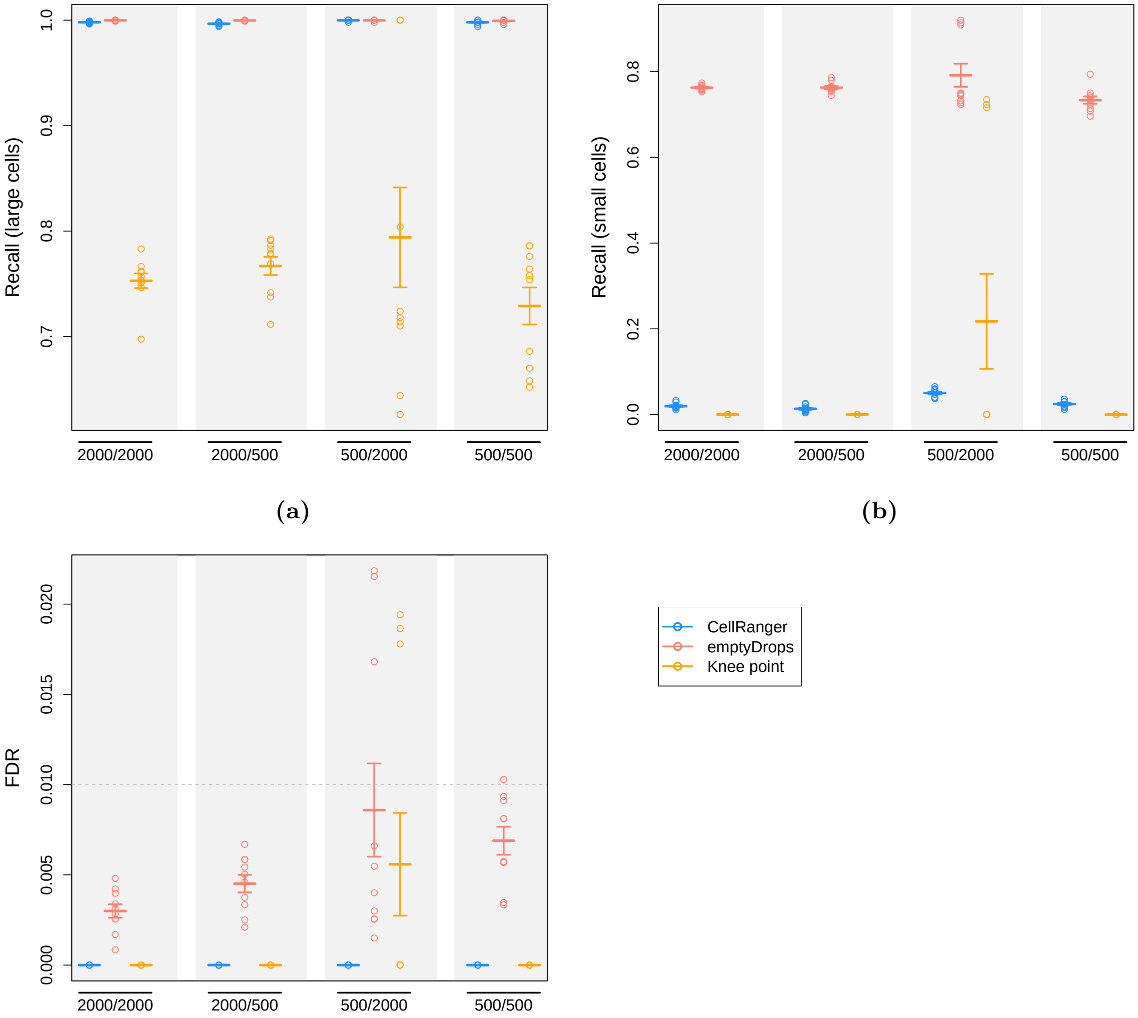


(a)

(b)

(c)

Supplementary Figure 3: Results for simulations based on the 293T cell line dataset, using three methods for detecting cell-containing droplets. Simulation scenarios are labelled as G_1/G_2 where G_1 and G_2 are the number of barcodes in the group of large and small cells, respectively. The recall for each group is shown as a proportion of the group size (a, b), and the FDR is calculated as the proportion of detected droplets that are empty (c). Each point represents the result of one simulation iteration, while the bar represents the mean across 10 iterations and the error bars represent the standard error of the mean. The dotted line represents the nominal FDR threshold (1%) for EmptyDrops.

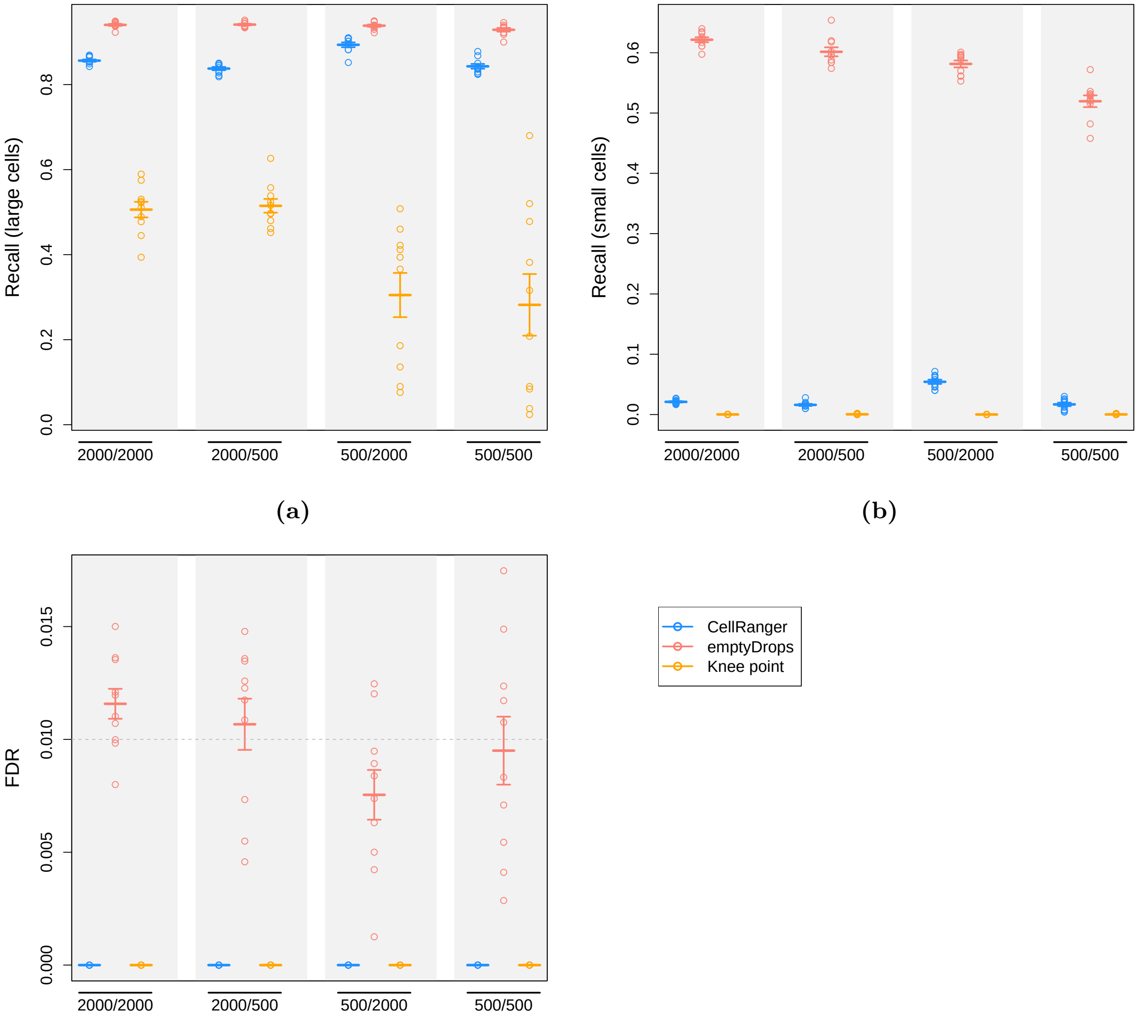


(a)

(b)

(c)

Supplementary Figure 4: Results for simulations based on the Jurkat cell line dataset, using three methods for detecting cell-containing droplets. Simulation scenarios are labelled as G_1/G_2 where G_1 and G_2 are the number of barcodes in the group of large and small cells, respectively. The recall for each group is shown as a proportion of the group size (a, b), and the FDR is calculated as the proportion of detected droplets that are empty (c). Each point represents the result of one simulation iteration, while the bar represents the mean across 10 iterations and the error bars represent the standard error of the mean. The dotted line represents the nominal FDR threshold (1%) for EmptyDrops.

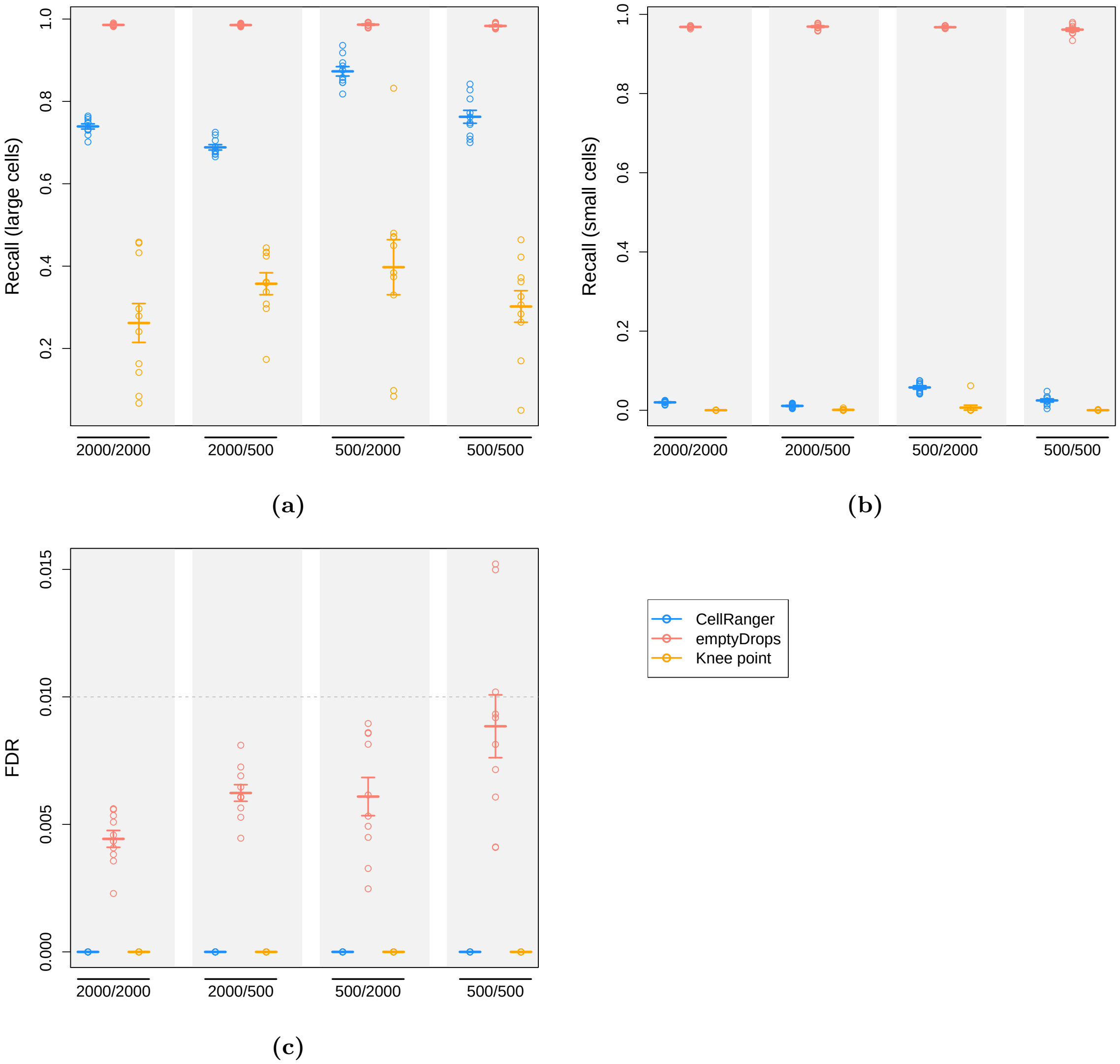


(a)

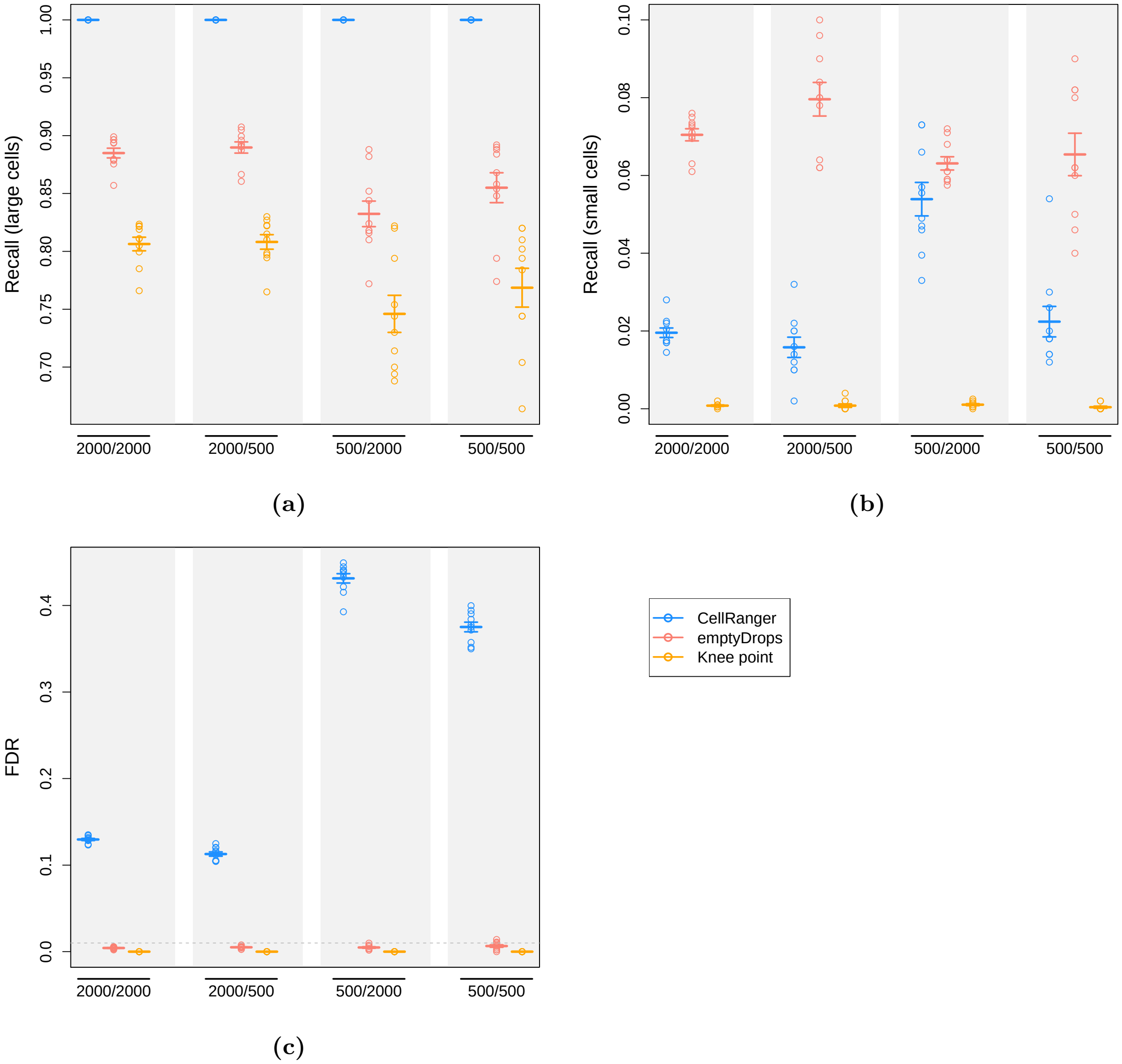
(b)

(c)

Supplementary Figure 5: Results for simulations based on the 9K brain cell dataset, using three methods for detecting cell-containing droplets. Simulation scenarios are labelled as G_1/G_2 where G_1 and G_2 are the number of barcodes in the group of large and small cells, respectively. The recall for each group is shown as a proportion of the group size (a, b), and the FDR is calculated as the proportion of detected droplets that are empty (c). Each point represents the result of one simulation iteration, while the bar represents the mean across 10 iterations and the error bars represent the standard error of the mean. The dotted line represents the nominal FDR threshold (1%) for EmptyDrops.



Supplementary Figure 6: Results for simulations based on the 900 brain cell dataset, using three methods for detecting cell-containing droplets. Simulation scenarios are labelled as G_1/G_2 where G_1 and G_2 are the number of barcodes in the group of large and small cells, respectively. The recall for each group is shown as a proportion of the group size (a, b), and the FDR is calculated as the proportion of detected droplets that are empty (c). Each point represents the result of one simulation iteration, while the bar represents the mean across 10 iterations and the error bars represent the standard error of the mean. The dotted line represents the nominal FDR threshold (1%) for EmptyDrops.

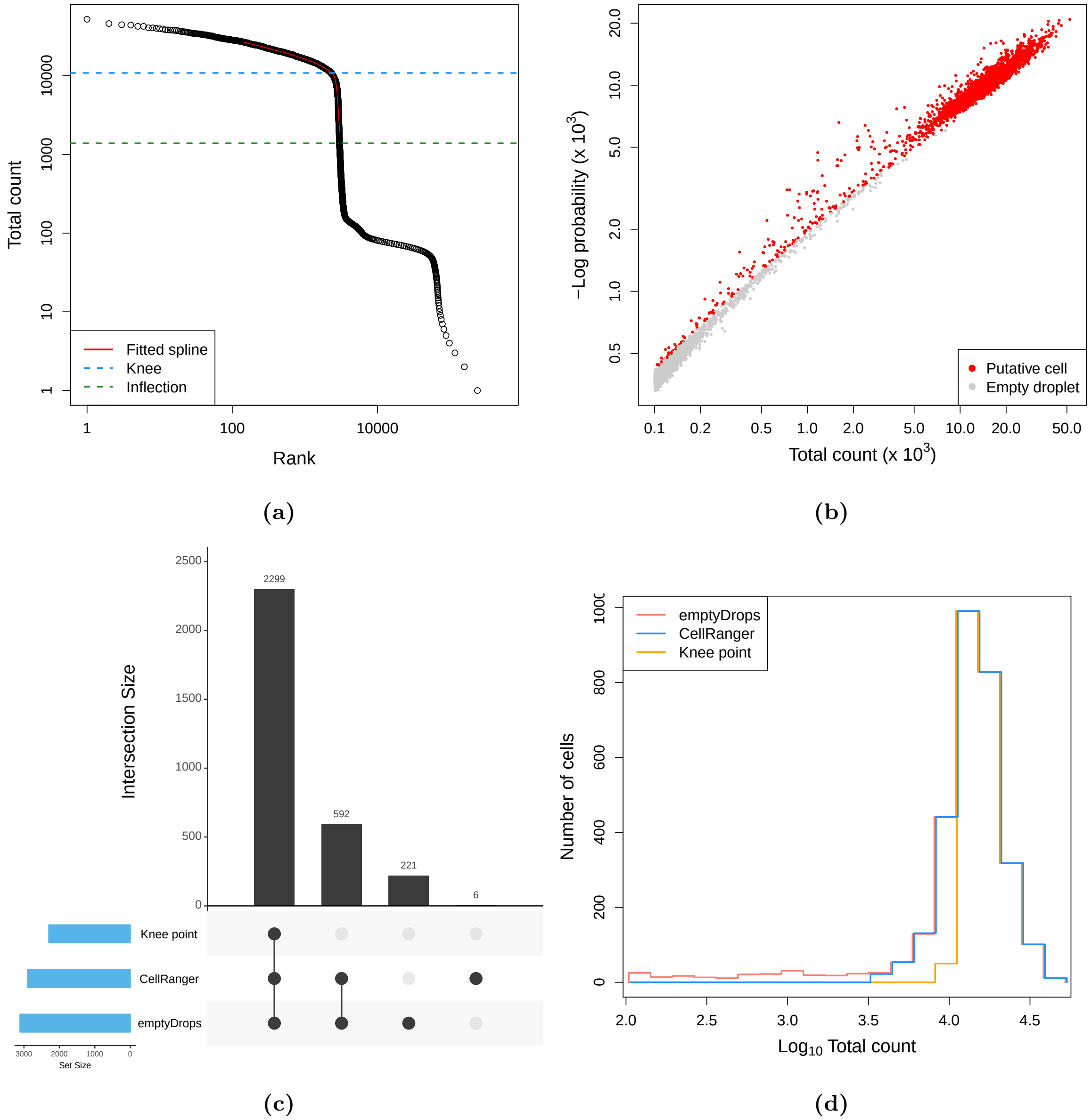


(a)

(b)

(c)

Supplementary Figure 7: Results for simulations based on the pan T cell dataset, using three methods for detecting cell-containing droplets. Simulation scenarios are labelled as G_1/G_2 where G_1 and G_2 are the number of barcodes in the group of large and small cells, respectively. The recall for each group is shown as a proportion of the group size (a, b), and the FDR is calculated as the proportion of detected droplets that are empty (c). Each point represents the result of one simulation iteration, while the bar represents the mean across 10 iterations and the error bars represent the standard error of the mean. The dotted line represents the nominal FDR threshold (1%) for EmptyDrops.



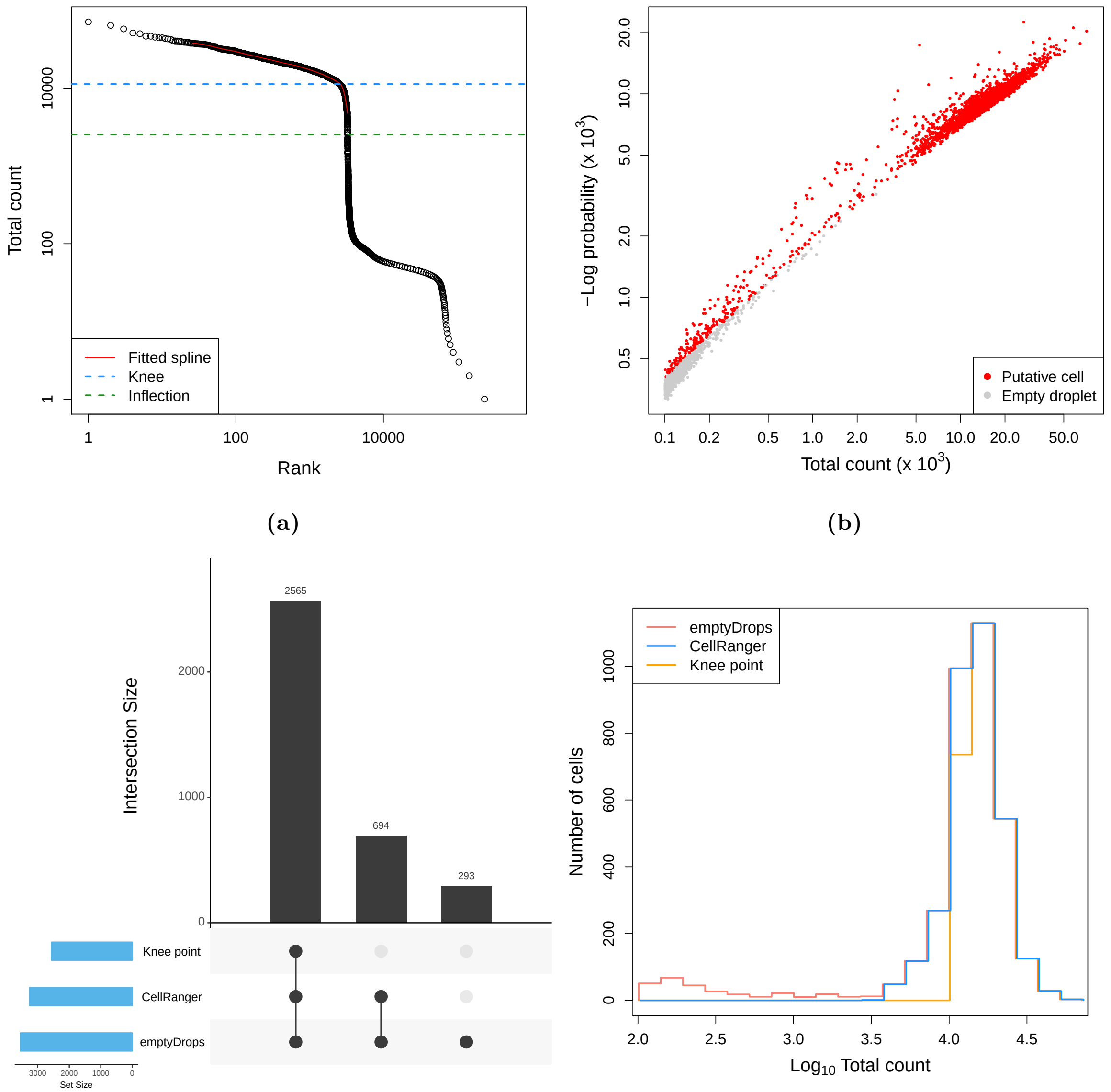
(a)

(b)

(c)

(d)

Supplementary Figure 8: Results of applying EmptyDrops and the other cell detection methods to the 293T cell line dataset. (a) A barcode rank plot showing the fitted spline used for knee point detection in EmptyDrops. The detected knee and inflection points are also shown. (b) The negative log-probability for each barcode in the multinomial model of EmptyDrops, plotted against the total count. Barcodes detected as putative cell-containing droplets at a FDR of 1% are labelled in red. Only barcodes with $t_b > T$ are shown. (c) An UpSet plot of the barcodes detected by each combination of methods (vertical bars). Horizontal bars represent the number of barcodes detected by each method. (d) Histogram outlines of the log-total count for barcodes detected by each method.



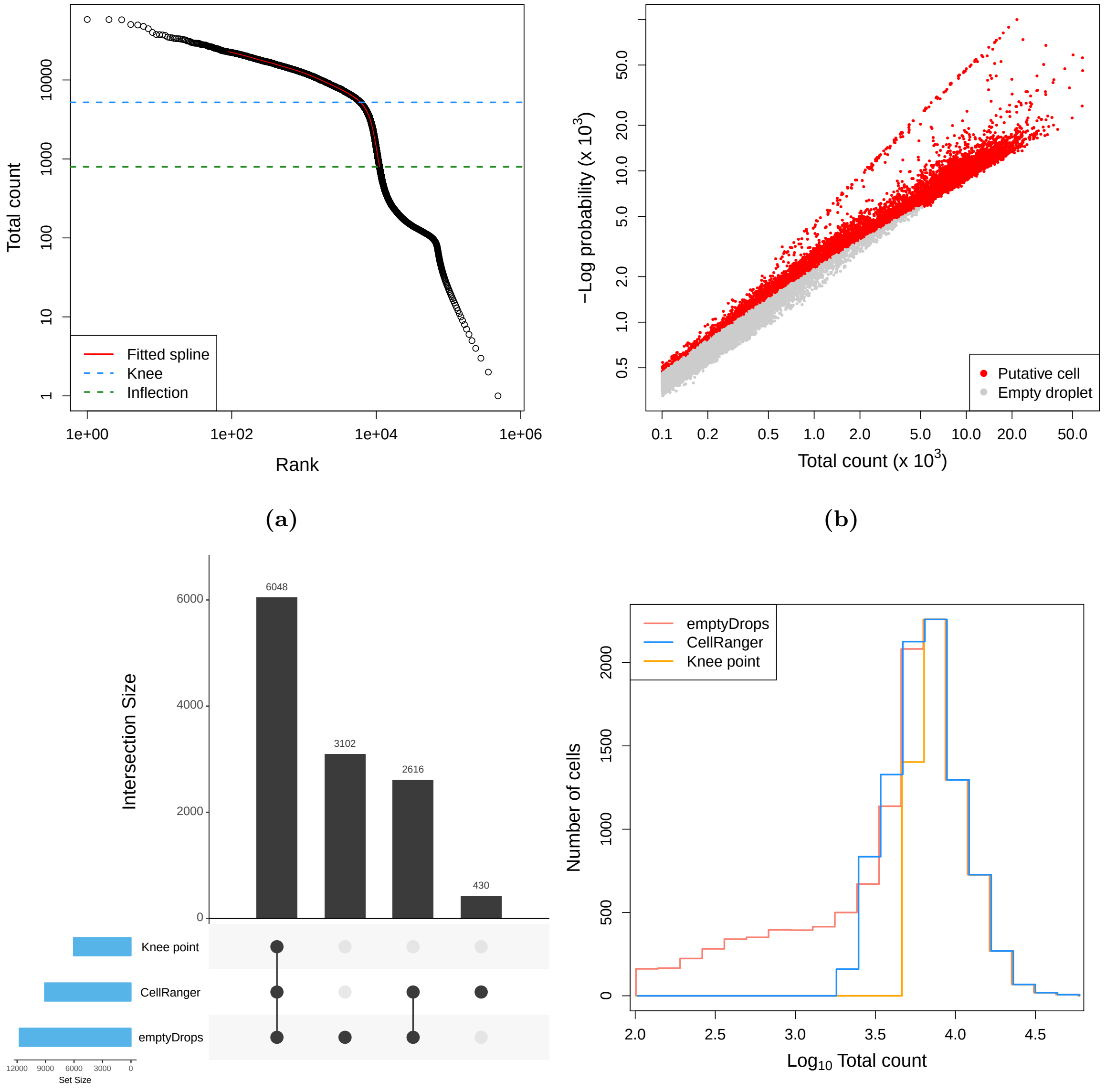
(a)

(b)

(c)

(d)

Supplementary Figure 9: Results of applying EmptyDrops and the other cell detection methods to the Jurkat cell line dataset. (a) A barcode rank plot showing the fitted spline used for knee point detection in EmptyDrops. The detected knee and inflection points are also shown. (b) The negative log-probability for each barcode in the multinomial model of EmptyDrops, plotted against the total count. Barcodes detected as putative cell-containing droplets at a FDR of 1% are labelled in red. Only barcodes with $t_b > T$ are shown. (c) An UpSet plot of the barcodes detected by each combination of methods (vertical bars). Horizontal bars represent the number of barcodes detected by each method. (d) Histogram outlines of the log-total count for barcodes detected by each method.



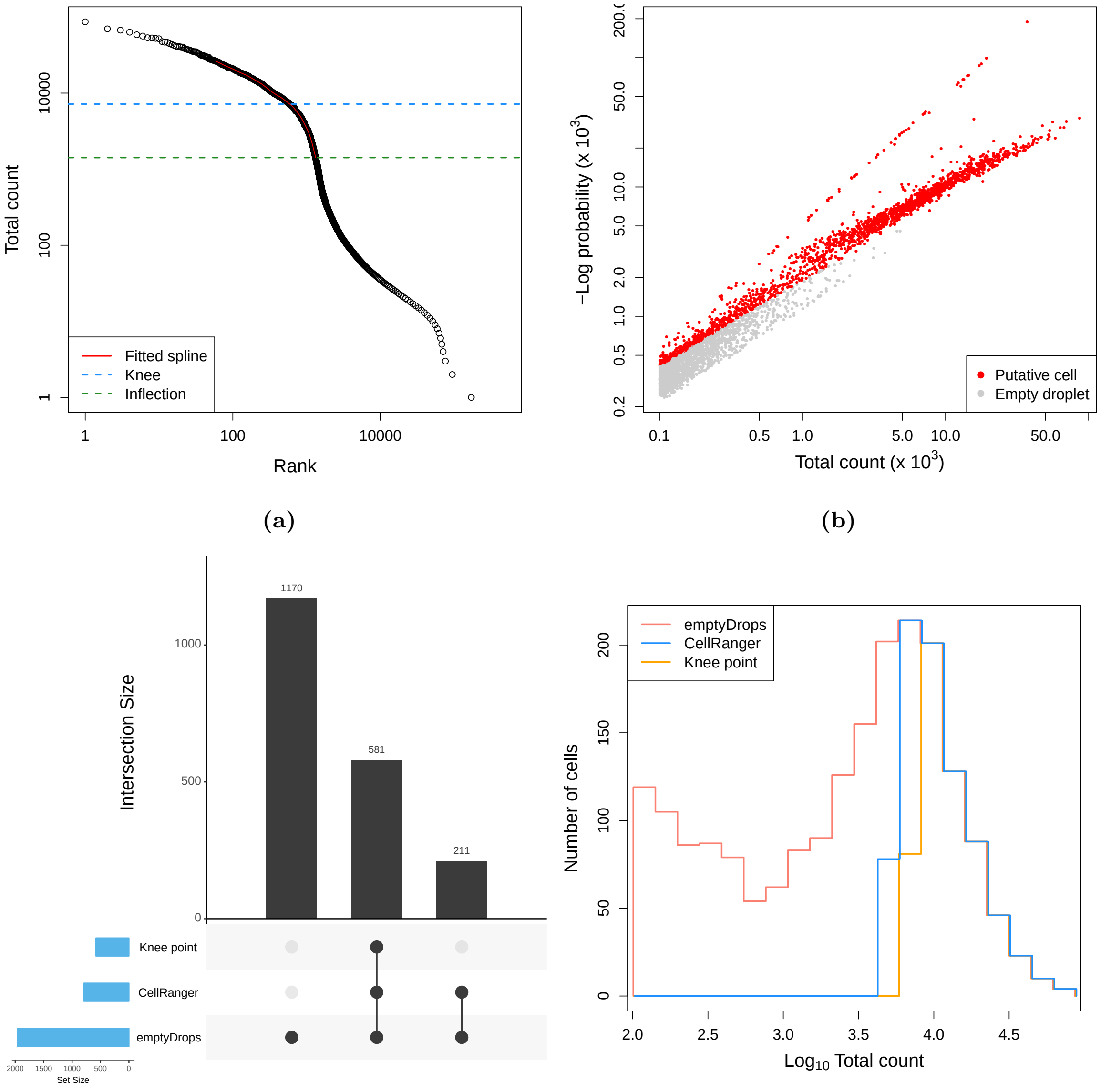
(a)

(b)

(c)

(d)

Supplementary Figure 10: Results of applying EmptyDrops and the other cell detection methods to the 9K brain cell dataset. (a) A barcode rank plot showing the fitted spline used for knee point detection in EmptyDrops. The detected knee and inflection points are also shown. (b) The negative log-probability for each barcode in the multinomial model of EmptyDrops, plotted against the total count. Barcodes detected as putative cell-containing droplets at a FDR of 1% are labelled in red. Only barcodes with $t_b > T$ are shown. (c) An UpSet plot of the barcodes detected by each combination of methods (vertical bars). Horizontal bars represent the number of barcodes detected by each method. (d) Histogram outlines of the log-total count for barcodes detected by each method.



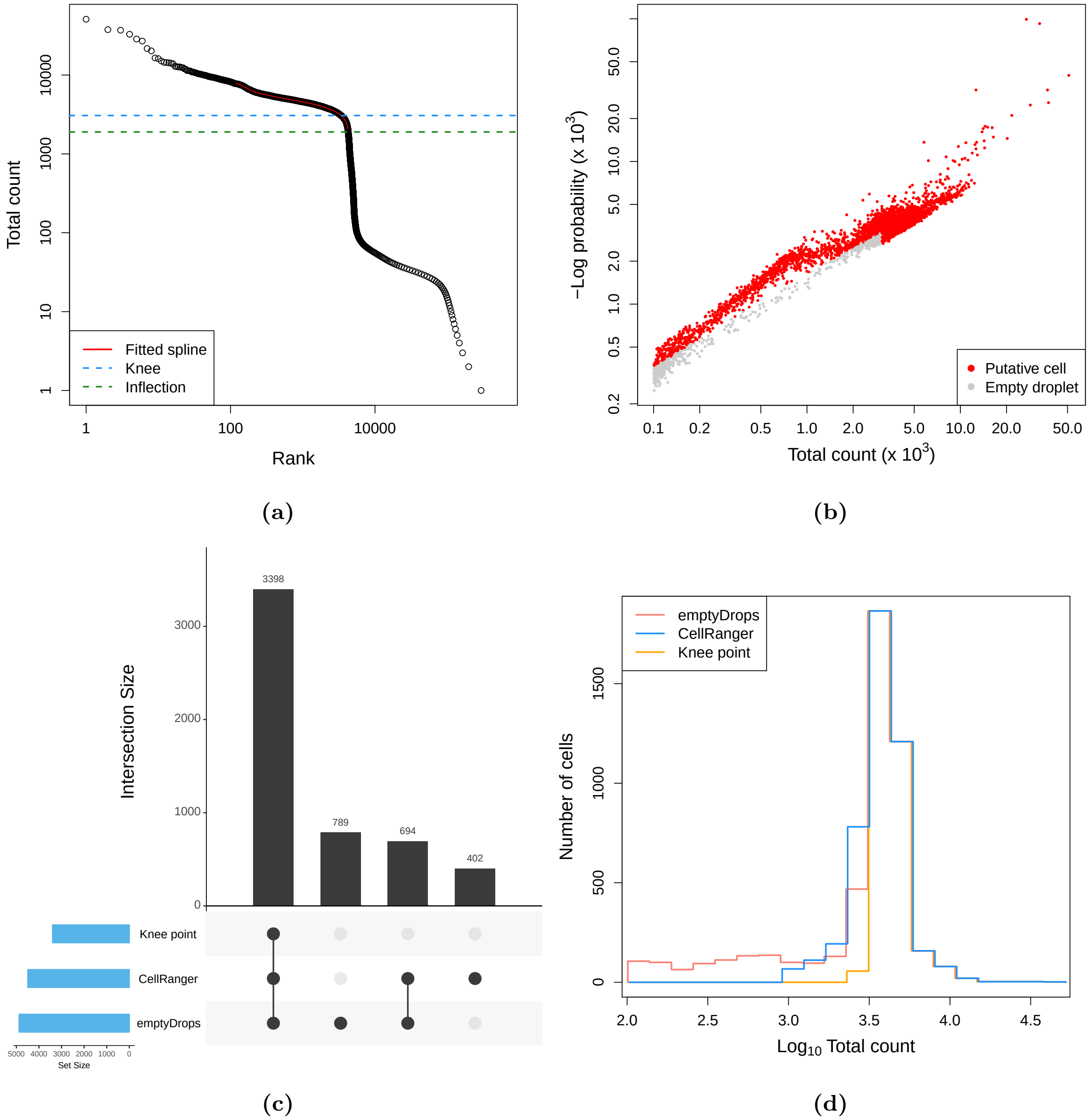
(a)

(b)

(c)

(d)

Supplementary Figure 11: Results of applying EmptyDrops and the other cell detection methods to the 900 brain cell dataset. (a) A barcode rank plot showing the fitted spline used for knee point detection in EmptyDrops. The detected knee and inflection points are also shown. (b) The negative log-probability for each barcode in the multinomial model of EmptyDrops, plotted against the total count. Barcodes detected as putative cell-containing droplets at a FDR of 1% are labelled in red. Only barcodes with $t_b > T$ are shown. (c) An UpSet plot of the barcodes detected by each combination of methods (vertical bars). Horizontal bars represent the number of barcodes detected by each method. (d) Histogram outlines of the log-total count for barcodes detected by each method.



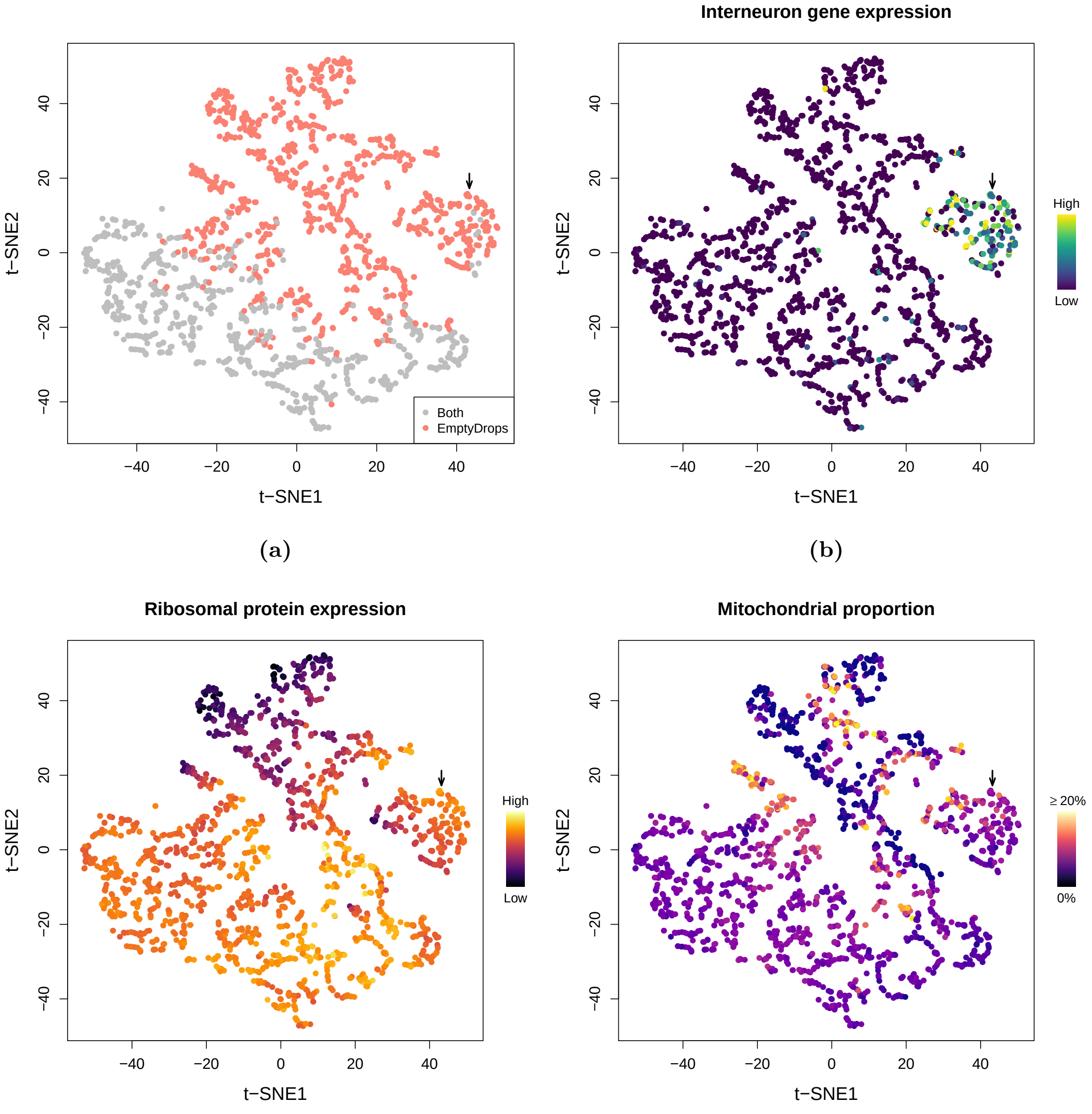
(a)

(b)

(c)

(d)

Supplementary Figure 12: Results of applying EmptyDrops and the other cell detection methods to the pan T cell dataset. (a) A barcode rank plot showing the fitted spline used for knee point detection in EmptyDrops. The detected knee and inflection points are also shown. (b) The negative log-probability for each barcode in the multinomial model of EmptyDrops, plotted against the total count. Barcodes detected as putative cell-containing droplets at a FDR of 1% are labelled in red. Only barcodes with $t_b > T$ are shown. (c) An UpSet plot of the barcodes detected by each combination of methods (vertical bars). Horizontal bars represent the number of barcodes detected by each method. (d) Histogram outlines of the log-total count for barcodes detected by each method.



(a)

(b)

(c)

(d)

Supplementary Figure 13: *t*-SNE plots for the 900 brain cell dataset, constructed using the first 100 principal components of the normalized log-expression matrix. Each point represents a cell that is coloured by (a) detection with EmptyDrops alone or both CellRanger and EmptyDrops, (b) expression of interneuron markers *Gad1*, *Gad2* and *Sla6c1*, (c) expression of ribosomal protein genes or (d) the proportion of counts assigned to mitochondrial genes. Expression in each cell was quantified as the sum of the normalized log-expression values across all genes in the relevant set. Mitochondrial proportions are capped at 20% to improve visibility. The arrow marks the putative interneuron population.