bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data

Walter Muskovic[1] and Joseph E Powell[1,2]

1. Garvan Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, The Kinghorn Cancer Centre, Darlinghurst, NSW 2010, Australia

2. UNSW Cellular Genomics Futures Institute, University of New South Wales, Sydney, NSW 2052, Australia

Walter Muskovic – w.muskovic@garvan.org.au
Joseph E Powell – j.powell@garvan.org.au

Abstract

Advances in droplet-based single cell RNA-sequencing (scRNA-seq) have dramatically increased throughput, allowing tens of thousands of cells to be routinely sequenced in a single experiment. In addition to cells, droplets capture cell-free "ambient" RNA predominately caused by lysis of cells during sample preparation. Samples with high ambient RNA concentration can create challenges in accurately distinguishing cell-containing droplets and droplets containing ambient RNA. Current methods to separate these groups often retain a significant number of droplets that do not contain cells – so called empty droplets. Additional to the challenge of identifying empty drops, there are currently no methods available to detect droplets containing damaged cells, which comprise of partially lysed cells – the original source of the ambient RNA. Here we describe *DropletQC*, a new method that is able to detect empty droplets, damaged, and intact cells, and accurately distinguish from one another. This approach is based on a novel quality control metric, the nuclear fraction, which quantifies for each droplet the fraction of RNA originating from unspliced, nuclear pre-mRNA. We demonstrate how *DropletQC* provides a powerful extension to existing computational methods for identifying empty droplets such as *EmptyDrops*. We have

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

implemented *DropletQC* as an *R* package, which can be easily integrated into existing single cell analysis workflows.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Main text

Droplet-based single cell RNA-sequencing (scRNA-seq) methods utilise microfluidics to encapsulate individual cells in nanolitre droplet emulsions, a technique that has dramatically increased throughput compared to plate-based protocols[1]. While encapsulating cells, droplets also capture cell-free ambient RNA, a complex mixture of transcripts released from damaged, stressed, and dying cells, often exacerbated during dissociation of solid tissues. This ambient RNA creates challenges for downstream analyses and the biological interpretation of results as most analysis methods are based on the assumption that a droplet contains RNA from a single cell. To combat this problem, several methods have been developed to estimate and remove its contribution to gene expression [2–4].

High levels of ambient RNA also create challenges in accurately identifying cell-containing droplets. This is a particular problem for data generated from solid tissues, where more fragile cells are more likely to become damaged during dissociation, and contribute to ambient RNA. We thus have three scenarios that need to be differentiated: empty droplets containing high concentrations of ambient RNA; droplets containing damaged cells; and droplets containing cells with limited ambient RNA. Using cut-offs based on the total number of RNA fragments assigned to each droplet, such as those originally proposed by Macosko *et al.* [5] and Zheng *et al.* [6], risks both including empty droplets and excluding small cells with below-average RNA content. The *EmptyDrops* method [7] addresses this issue through a more sophisticated approach, calculating the profile of the ambient RNA pool and testing each barcode for significant deviations from this profile. A favoured alternative to simple UMI cut-offs, this method has been integrated as the default cell-calling algorithm in the widely used *CellRanger* pipeline [6]. However, cell-free droplets with high ambient RNA concentration and damaged cells are still retained by this method.

Here, we present *DropletQC*, a new method that is able to simultaneously improve the detection of cell free droplets and droplets containing damaged cells. Taking advantage of the observation that unspliced and spliced mRNAs can be distinguished in common scRNA-seq protocols [8], we develop a novel metric: the nuclear fraction. The nuclear fraction quantifies, for each droplet, the proportion of RNA originating from unspliced pre-mRNA. Ambient RNA consists predominantly of mature cytoplasmic mRNA. This may arise as RNA is released from damaged cells in which the nuclear envelope remains intact, or capped and

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

polyadenylated transcripts may be more stable in the extracellular environment (**Figure 1**). Regardless, droplets that contain only ambient RNA have a low nuclear fraction compared to droplets containing cells. In contrast, damaged cells due to the depletion of cytoplasmic RNA, will have a higher nuclear fraction compared to intact cells. By using the nuclear fraction score in combination with UMIs per droplet, we are able to accurately distinguish between empty droplets, damaged cells, and intact cells.

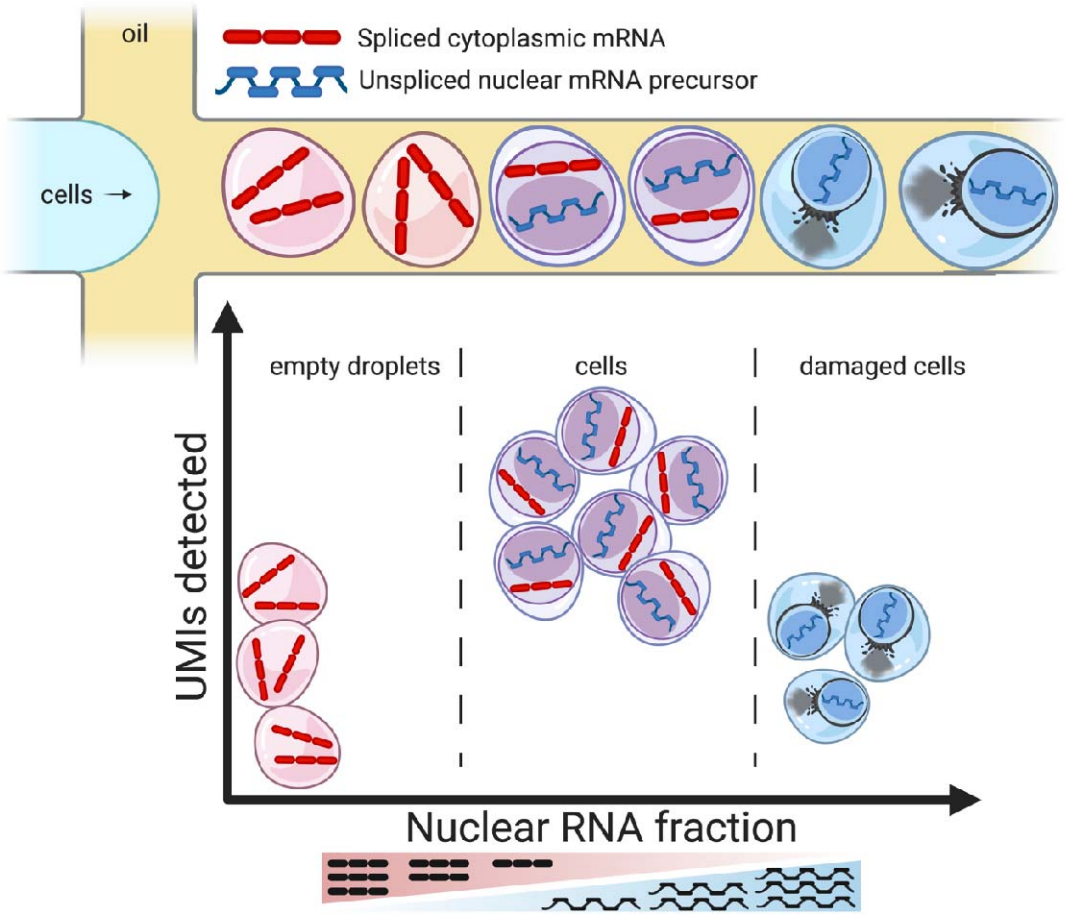


Figure 1 | Illustration of how the nuclear fraction, in combination with the library size of each droplet, can be used to separate the populations of empty droplets, intact cells and damaged cells.

To assess the ability of *DropletQC* to identify both empty droplets and droplets containing damaged cells, we applied it to four independent scRNA-seq datasets; embryonic mouse brain, glioblastoma tumour, peripheral blood mononuclear cells (PBMCs), and Hodgkin's lymphoma tumour. To determine whether *DropletQC* could identify empty droplets missed by current methods, all barcodes were first filtered using *EmptyDrops*, as implemented in *DropletUtils* [7]. *DropletQC* identified an additional 9.5% of mouse brain, 6.0% of

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Hodgkin's lymphoma, 4.0% of glioblastoma and 0.4% of PBMCs as empty droplets (**Figure 2**, **Table S1**). Cells from dissociated tissue (**Figure 2a-c**) contained more empty droplets with high RNA content than PBMCs (**Figure 2d**), suggesting ambient RNA may be released from cells damaged during sample preparation.

Following identification of empty droplets, droplets containing damaged cells are identified using expectation maximisation and a Gaussian mixture model to separate them from droplets containing intact cells. As both the total UMI count and nuclear fraction scores display distinct distributions for different cell types (**Figure S1**), it is necessary to first group cells by type. Cells were annotated for each sample using a combination of gene markers and supervised classification with *scPred* [9]. Of the remaining cells, 14.0% of mouse brain, 5.2% of Hodgkin's lymphoma, 9.8% of glioblastoma tumour cells and one PBMC cell were identified as damaged cells (**Table S1**).

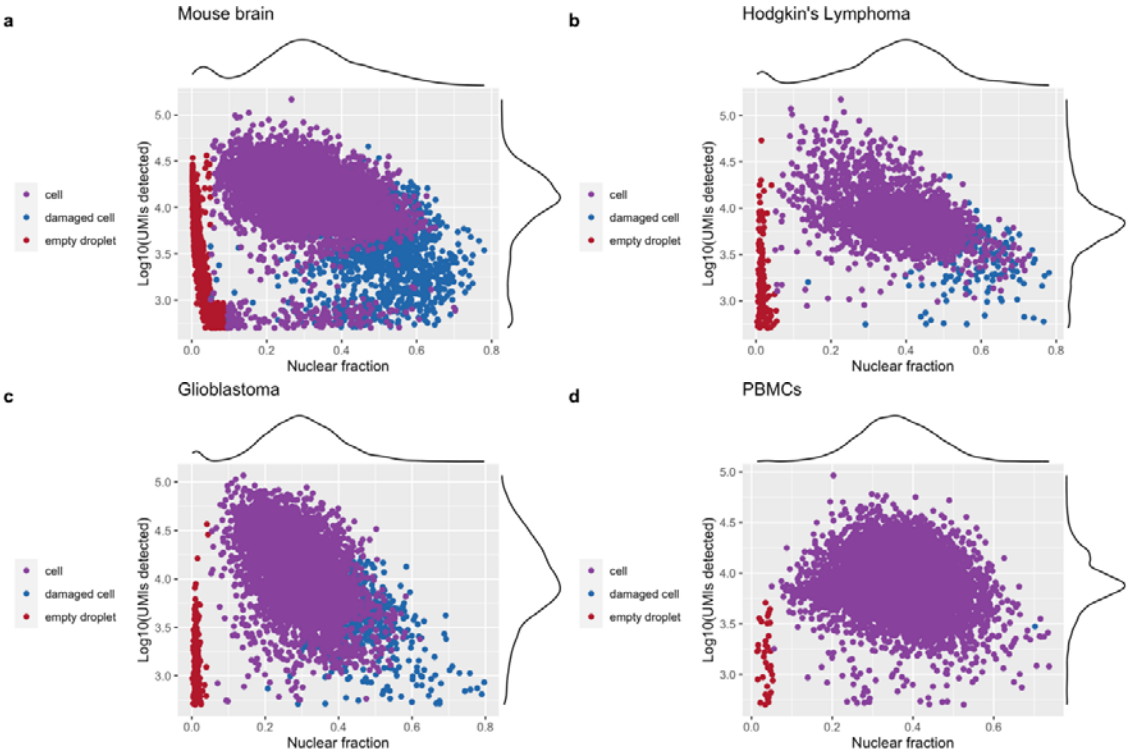


Figure 2 | DropletQC identifies empty droplets and damaged cells in four heterogeneous scRNA-seq datasets. Total UMI counts (y-axis) and nuclear fraction scores (x-axis) are shown for each cell, with colours representing the status of each cell assigned by DropletQC. Empty droplets contain less RNA than cells and a higher fraction of cytoplasmic RNA (low nuclear fraction score). Damaged cells contain less RNA than intact cells and a higher proportion of unspliced RNA fragments (high nuclear fraction score).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

As an additional test of the ability of *DropletQC* to identify damaged cells, we applied the method to data from a recent investigation on the effects of cryopreservation on the transcriptomes of macaque microglia [10]. *DropletQC* revealed an increase in the proportion of damaged cells following cryopreservation from 4.1% to 13.8% (**Figure S2, Table S2**). These findings have implications for the suitability of prospectively archiving samples for scRNA-seq studies through cryopreservation and demonstrates the utility of *DropletQC* for similar studies.

By default, the *DropletQC* software provides a flag for empty droplets and damaged cells, but does not automatically remove them from the dataset. Depending on the biological analyses, damaged cells may retain useful information, and as such it may be desirable to retain this metadata throughout downstream analyses. Similarly, cells such as erythrocytes, which contain small amounts of mature mRNA, may be misidentified as empty droplets and can be rescued downstream if desired.

For samples with large percentages of ambient RNA, some damaged cells and empty droplets may be missed by *DropletQC*. However, these can be identified by their low RNA content (**Figure 2a**) and may be easily flagged using a minimum UMI threshold. Calculation of the nuclear fraction, identification of empty droplets and damaged cells are implemented as separate functions within the *DropletQC* package. In summary, we have shown that *DropletQC* is able to successfully identify both empty droplets and damaged cells in data from a range of tissue types.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Nuclear fraction calculation

The *DropletQC* method first calculates the nuclear fraction for each droplet, which is the proportion of RNA fragments that originate from intronic regions. It is calculated as:

$$NF_i = \frac{\Sigma(IR_i)}{\Sigma(IR_i + ER_i)}$$

Where NF_i is the nuclear fraction for droplet i, IR_i are the reads that map to intronic regions for droplet i, and ER_i are those that map to exonic regions. We have implemented two methods to map reads to either intronic or exonic regions. The first, takes advantage of region tags, such as those added by 10x Genomics' Cell Ranger count analysis pipeline that identify the region type of each genome-aligned RNA fragment; exonic, intronic or intergenic. These are efficiently counted using the nuclear_fraction_tags function to calculate a nuclear fraction score for each provided cell barcode. Alternatively, if region tags are missing, our second method assesses aligned reads for overlap with intronic regions using the nuclear_fraction_annotation function in combination with a user-provided gene annotation file. To speed up processing of indexed, coordinate-sorted alignment files, reads are split across a user-specified number of genomic regions to allow parallel computation. The four samples presented in the manuscript were processed with 8 CPUs and 16Gb of RAM with an average processing time of 106 seconds per 100 million reads using the nuclear_fraction_tags function and 132 seconds per 100 million reads using the nuclear_fraction_annotation function.

Identifying empty droplets and damaged cells

Empty droplets are classified as all barcodes that fall below a defined nuclear fraction threshold. To identify a suitable threshold, a kernel density estimate is calculated using the nuclear fraction scores. The first derivative of the estimate is then calculated to identify the local minimum immediately following the first peak, corresponding to the population of empty droplets. If the automatically selected cut-off misidentifies the empty droplet population, two user-adjustable parameters are provided; a nuclear fraction threshold and a total UMI threshold, above which all barcodes are marked as cells.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To identify droplets containing damaged cells, barcodes are assessed separately for each cell type. It is assumed damaged cells have both a lower UMI count and higher nuclear fraction score than undamaged cells. We therefore use a two component (*k*) gaussian mixture model to classify droplets containing damaged cells:

$$P(X\,|\mu,\sigma,\alpha) = \alpha_1 N(X|\mu_1,\sigma_1^2) + \alpha_2 N(X|\mu_2,\sigma_2^2) \text{ (Eq. 1)}$$

Where *X* is a dataset with $log_{10}(UMI)$ and estimated nuclear fractions for *1-n* droplets of a given cell type. μ and $\sigma^2$are the mean and variance, and α represents the mixing weight of a given component. The initial model parameters are calculated as:

$$\mu_k = \frac{\sum_i^{N_k} x_{i,k}}{N_k}$$

$$\sigma_k^2 = \frac{\sum_i^{N_k} (x_{i,k} - \mu_k)^2}{N_k}$$

$$\alpha_k = \frac{N_k}{N}$$

Where $N_k$is the number of data points in the kth component. Following the initialisation, we estimate parameters using expectation maximisation by asking what is the posterior probability that a droplet (x_i) belongs to component k_j:

$$P\left(x_i \in k_j \middle| x_i\right) = \frac{P\left(x_i \middle| x_i \in k_j\right) P(k_j)}{P(x_i)} \text{ (Eq. 2)}$$

Where,

$$P\left(x_i \middle| x_i \in k_j\right) = N\left(x_i \middle| \mu_{kj}, \sigma_{kj}^2\right)$$

$$P\left(k_j\right) = \alpha_{kj}$$

$$P(x_i) = \sum_{k=1}^{K} \alpha_k N(x_i|\mu_k, \sigma_k^2)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$N_k$in the initial component parameters are replaced with the posterior probability and recalculated, with these steps repeated until convergence determined using the Bayesian information criterion. This model identifies the minimum separation required between the identified distributions for a population of droplets to be marked as damaged. We then label droplets as containing a damaged cell based on: a higher mean nuclear fraction and lower mean UMI than the cell population; a mean nuclear fraction greater than the cell population mean by a user-adjustable amount (default 0.15); a mean UMI count lower than the cell population (default 50%).

Data

Cell filtering and annotation

Prior to calculating the nuclear fraction score, all cell barcodes were assessed for a significant deviation from the ambient RNA expression pattern using the *EmptyDrops* method implemented in the *DropletUtils* [7]. The lower bound on the total UMI count used to identify empty droplets was increased from 100 to 500 and all other parameters were left at their default values. Barcodes below a false discovery rate threshold of 1% were excluded. Remaining barcodes were additionally filtered for a maximum mitochondrial gene content of 15% to exclude low quality cells. Mouse brain and PBMC cell types were annotated by supervised classification with the *scPred* [9] using annotated PBMC [11], mouse brain [12] and developing mouse brain [13] reference datasets. The glioblastoma sample cell types were identified using cell-type specific gene markers for oligodendrocytes (*MAG*, *MOG*, *MBP*), microglia/macrophages (*C1QA*, *AIF1*, *LAPTM5*), T cells (*CD2*, *CD3D*, *CD3E*) and endothelial cells (*CD34*, *ESAM*, *APOLD1*) [14–17]. Hodgkin's lymphoma cell types were classified using marker genes for B cells (*MS4A1*), macrophages (*CD68*, *IDO1*), plasmacytoid dendritic cells (*CLEC4C*, *NRP1*), erythrocytes (*HBB*, *HBA1*, *HBA2*), cytotoxic T cells (*GZMA*, *GZMK*, *IFNG*), regulatory T cells (*FOXP3*, *IL2RA*, *IKZF2*), T helper cells (*CXCL13*, *PDCD1*, *FABP5*), naïve T cells (*CCR7*, *IL7R*, *LEF1*), progenitor (*CD34*) and mast cells (*TPSAB1*, *TPSB2*, *KIT*) [18,19].

Availability of data and materials

The four single-cell gene expression datasets presented in the manuscript are made publicly available through the 10x Genomics website: https://support.10xgenomics.com/single-cell-gene-expression/datasets. The macaque microglia expression data is available from the NCBI GEO database, under accession GSE162663. All of the code used to produce the analyses

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and figures presented in the manuscript, along with links to individual datasets, are available through GitHub at https://github.com/powellgenomicslab/dropletQC_paper. *DropletQC* is available as an *R* package at https://github.com/powellgenomicslab/DropletQC.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade [Internet]. Nat. Protoc. Nature Publishing Group; 2018 [cited 2021 Mar 30]. p. 599–604. Available from: https://www.fluidigm.com/about/aboutfluidigm

2. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol [Internet]. BioMed Central Ltd.; 2020 [cited 2021 Feb 22];21:57. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1950-6

3. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience [Internet]. NLM (Medline); 2020 [cited 2021 Feb 22];9:1–10. Available from: http://orcid.org/0000-0003-0937-5290

4. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat Methods [Internet]. Nature Research; 2020 [cited 2021 Feb 22];17:615–20. Available from: https://doi.org/10.1038/s41592-020-0820-1

5. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell [Internet]. Cell Press; 2015 [cited 2021 Feb 22];161:1202–14. Available from: http://dx.doi.org/10.1016/j.cell.2015.05.002

6. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun [Internet]. Nature Publishing Group; 2017 [cited 2021 Feb 22];8:1–12. Available from: www.nature.com/naturecommunications

7. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol [Internet]. BioMed Central Ltd.; 2019 [cited 2021 Feb 17];20:63. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y

8. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature [Internet]. Nature Publishing Group; 2018 [cited 2019 Nov 27];560:494–8. Available from: http://www.nature.com/articles/s41586-018-0414-6

9. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol [Internet]. 2019 [cited 2019 Dec 18];20:264. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1862-5

10. Morsey B, Niu M, Dyavar SR, Fletcher C V., Lamberty BG, Emanuel K, et al. Cryopreservation of microglia enables single-cell RNA sequencing with minimal effects on disease-related gene expression patterns. iScience. Elsevier BV; 2021;24:102357.

11. Senabouth A, Andersen S, Shi Q, Shi L, Jiang F, Zhang W, et al. Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. NAR Genomics Bioinforma [Internet]. Oxford University Press (OUP); 2020 [cited 2021 Feb 17];2. Available from: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa034/5836690

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

12. Yao Z, Nguyen TN, van Velthoven CTJ, Goldy J, Sedeno-Cortes AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. bioRxiv [Internet]. 2020;2020.03.30.015214. Available from: http://biorxiv.org/content/early/2020/03/31/2020.03.30.015214.abstract

13. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science (80-) [Internet]. American Association for the Advancement of Science; 2018 [cited 2021 Feb 17];360:176–82. Available from: http://science.sciencemag.org/

14. Dusart P, Hallström BM, Renné T, Odeberg J, Uhlén M, Butler LM. A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes. Cell Rep. Elsevier B.V.; 2019;29:1690-1706.e4.

15. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. n Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma.A. Cell [Internet]. Elsevier; 2019 [cited 2019 Oct 14];178:835-849.e21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31327527

16. Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nat Commun [Internet]. Nature Publishing Group; 2020 [cited 2020 Jul 10];11:3406. Available from: http://www.nature.com/articles/s41467-020-17186-5

17. Wang L, Babikir H, Müller S, Yagnik G, Shamardani K, Catalan F, et al. The Phenotypes of Proliferating Glioblastoma Cells Reside on a Single Axis of Variation. Cancer Discov [Internet]. 2019;9:1708 LP – 1719. Available from: http://cancerdiscovery.aacrjournals.org/content/9/12/1708.abstract

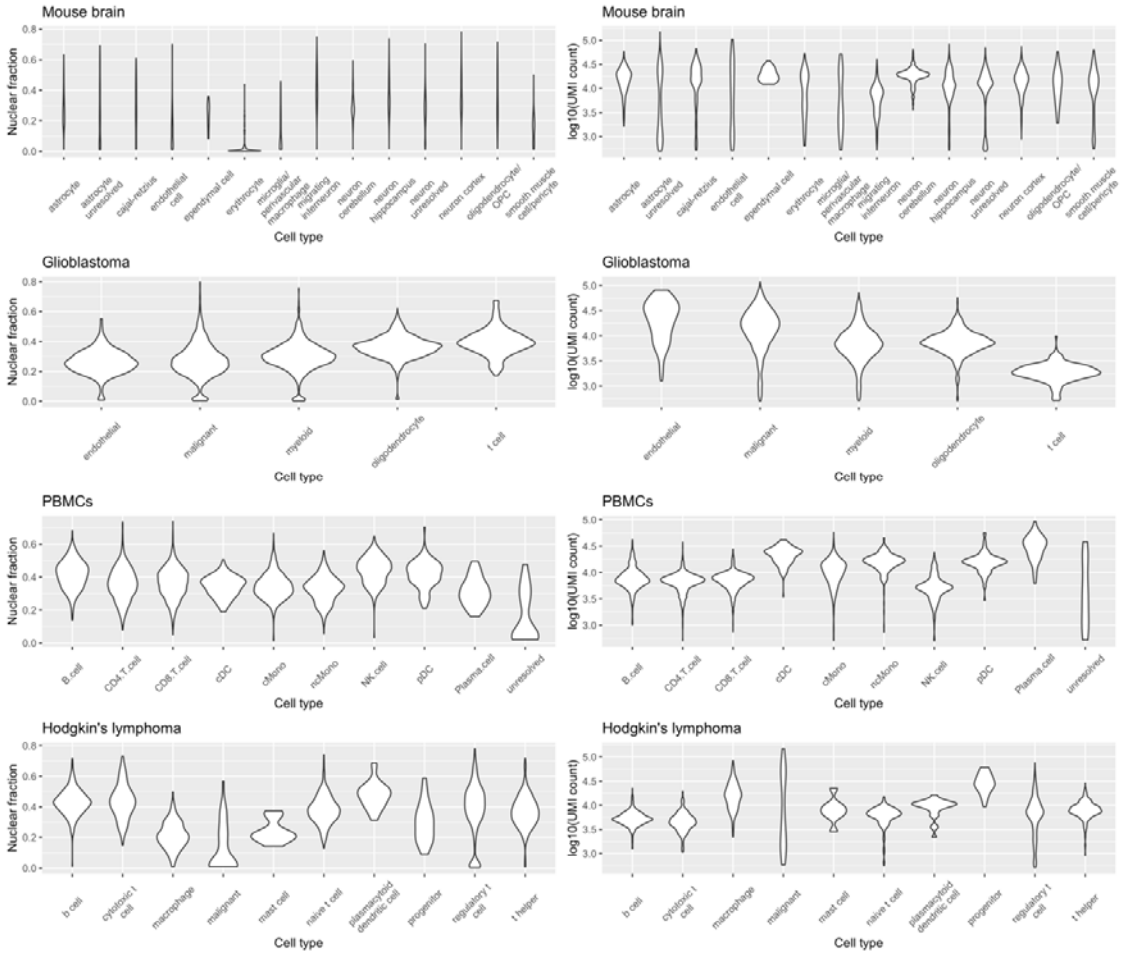
18. Aoki T, Chong LC, Takata K, Milne K, Hav M, Colombo A, et al. Single-cell transcriptome analysis reveals disease-defining t-cell subsets in the tumor microenvironment of classic hodgkin lymphoma. Cancer Discov [Internet]. American Association for Cancer Research Inc.; 2020 [cited 2021 Feb 17];10:406–21. Available from: http://cancerdiscovery.aacrjournals.org/

19. Schafflick D, Xu CA, Hartlehnert M, Cole M, Schulte-Mecklenbeck A, Lautwein T, et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. Nat Commun [Internet]. Nature Research; 2020 [cited 2021 Feb 17];11:1–14. Available from: https://doi.org/10.1038/s41467-019-14118-w

20. Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. Technische Universitaet Wien; 2016;8:289–317.

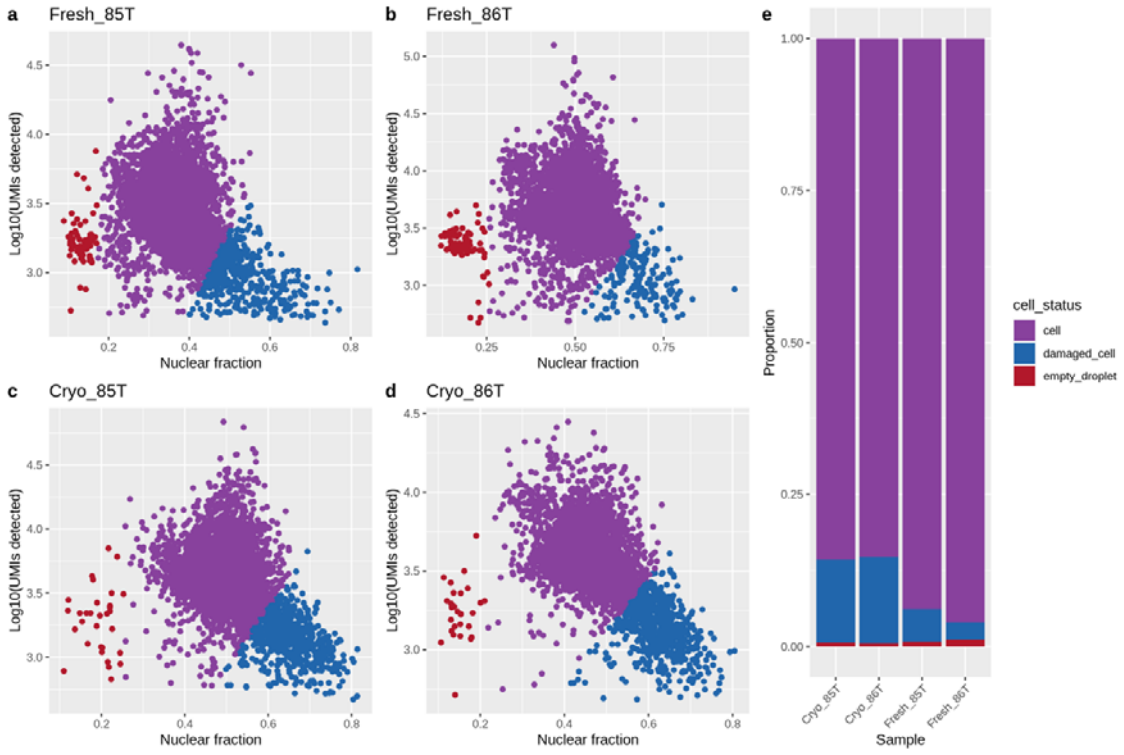
bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Figures



Supplementary Figure 1 | Different cell types have distinct distributions of nuclear fraction scores and UMI counts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Supplementary Figure 2 | DropletQC identifies an increased proportion of damaged cells in cryopreserved microglia samples. (c-d) compared to fresh tissue samples (a-b). Total UMI counts (y-axis) and nuclear fraction scores (x-axis) are shown for each cell, with colours representing the status of each cell assigned by DropletQC. The stacked bar chart (e) illustrates the proportion of empty droplets and damaged cells for each sample.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Tables

Sample	Empty droplets	Damaged cells
Embryonic mouse brain	9.53%	14.00%
Glioblastoma tumour	3.99%	9.77%
Peripheral blood mononuclear cells	0.37%	0.01%
Hodgkin's lymphoma tumour	6.04%	5.16%

Supplementary Table 1 | Summary of the percentage of empty droplets and damaged cells identified in four heterogeneous scRNA-seq datasets after filtering with EmptyDrops and a maximum mitochondrial gene content of 15%.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.02.454717; this version posted August 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Sample	Empty droplets	Damaged cells
Fresh_85T	0.80%	5.40%
Fresh_86T	1.15%	2.84%
Cryo_85T	0.64%	13.40%
Cryo_86T	0.57%	14.10%

Supplementary Table 2 | Summary of the percentage of empty droplets and damaged cells identified in four macaque microglia scRNA-seq datasets.

345