

METHOD

Open Access

CB2 improves power of cell detection in droplet-based single-cell RNA sequencing data



Zijian Ni¹, Shuyang Chen¹, Jared Brown¹ and Christina Kendziorski^{2*}

* Correspondence: c.kendziorski@gmail.com

²Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

Full list of author information is available at the end of the article

Abstract

An important challenge in pre-processing data from droplet-based single-cell RNA sequencing protocols is distinguishing barcodes associated with real cells from those binding background reads. Existing methods test barcodes individually and consequently do not leverage the strong cell-to-cell correlation present in most datasets. To improve cell detection, we introduce CB2, a cluster-based approach for distinguishing real cells from background barcodes. As demonstrated in simulated and case study datasets, CB2 has increased power for identifying real cells which allows for the identification of novel subpopulations and improves the precision of downstream analyses.

Keywords: Single-cell RNA-seq, Droplet-based protocols, Cell detection

Background

Droplet-based single-cell RNA sequencing (scRNA-seq) [1] is a powerful and widely used approach for profiling genome-wide gene expression in individual cells. Current commercial droplet-based technologies utilize gel beads [2], each containing oligonucleotide indexes made up of bead-specific barcodes combined with unique molecular identifiers (UMIs) [3] and oligo-dT tags to prime polyadenylated RNA. Single cells of interest are combined with reagents in one channel of a microfluidic chip, and gel beads in another, to form gel beads in emulsion, or GEMs. Oligonucleotide indexes bind polyadenylated RNA within each GEM reaction vesicle before gel beads are dissolved releasing the bound oligos into solution for reverse transcription. By design, each resulting cDNA molecule contains a UMI and a GEM-specific barcode. Indexed cDNA is pooled for PCR amplification and sequencing resulting in a data matrix of UMI counts for each barcode (Additional file 1: Figure S1).

Ideally, each barcode will tag mRNA from an individual cell, but this is often not the case in practice. In most datasets, more than 90% of GEMs do not contain viable cells, but rather contain ambient RNA excreted by cells in solution or as a product of cell lysis [2]. As a result, an important challenge in pre-processing droplet-based



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

scRNA-seq data is distinguishing those barcodes corresponding to real cells from those binding ambient, or background, RNA.

Early methods to address this challenge defined real cells as those barcodes with total read counts exceeding some threshold [1, 2]. Such methods are suboptimal as they discard small cells as well as those expressing relatively few genes. To address this, Lun et al. [4] developed EmptyDrops (ED), an approach to identify individual barcodes with distributions varying from a background distribution. Similar to previous approaches [1, 2], ED identifies an upper threshold and defines real cells as those barcodes with counts above the threshold. As a second step, ED uses all barcodes with counts below a lower threshold to estimate a background distribution of ambient RNA against which remaining barcodes are tested. Those having expression profiles significantly different from the background distribution are deemed real cells. The ED approach is current state-of-the-art in the field. However, given that ED performs tests for each barcode individually, it does not leverage the strong correlation observed between cells and, consequently, compromises power for identifying cells in many datasets.

To increase the power for identifying real cells, we propose CB2, a cluster-based approach for distinguishing real cells from background barcodes in droplet-based scRNA-seq experiments. CB2 extends the ED framework by introducing a clustering step that groups similar barcodes, then conducts a statistical test to identify groups with expression distributions that vary from the background (Fig. 1, Additional file 1: Figure S2). CB2 is implemented in the R package *scCB2*.

Results

CB2 was evaluated and compared with ED on simulated and case study data. In SIM IA, counts are generated as in Lun et al. [4]. Briefly, given an input dataset, an inflection point dividing low from high-count barcodes is determined. Low count barcodes are pooled to estimate the background distribution. Background barcodes are sampled from this distribution to match the total number and size of barcodes below the inflection point in the input dataset. Six thousand real cells are then generated as follows. First, 2000 barcodes are randomly sampled from the high-count barcodes (referred to as G_1 cells [4]); a second set of 2000 high-count barcodes is sampled and then downsampled by 90% to give G_2 cells; the third set ($G_{1.5}$) is obtained by sampling 2000 barcodes from the high-count range and downsampling by 50%. We note that in Lun et al. [4], only G_1 and G_2 cells were considered. Here, $G_{1.5}$ cells were added to better reflect real data. Additional file 1: Figure S3 shows increased power of CB2 with well controlled false discovery rate (FDR) for the 6 datasets considered in Lun et al. [4] as well as 4 additional datasets. SIM IB, also considered by Lun et al. [4], is similar to SIM IA, but in SIM IB 10% of the genes in the real cells are shuffled making the real cells more different from the background and therefore easier to identify (Additional file 1: Figure S4). Additional file 1: Figure S5 shows the increased power of CB2 is maintained.

To further evaluate CB2, we applied CB2 and ED to the ten case study datasets used to generate the simulated data as well as one additional dataset considered in the ED case study and compared the number of cells identified in common as well as those uniquely identified by each approach. Additional file 2: Table S1 shows that CB2 finds 24% more cells on average (range 4–81%). Of the extra cells identified, 88% on average

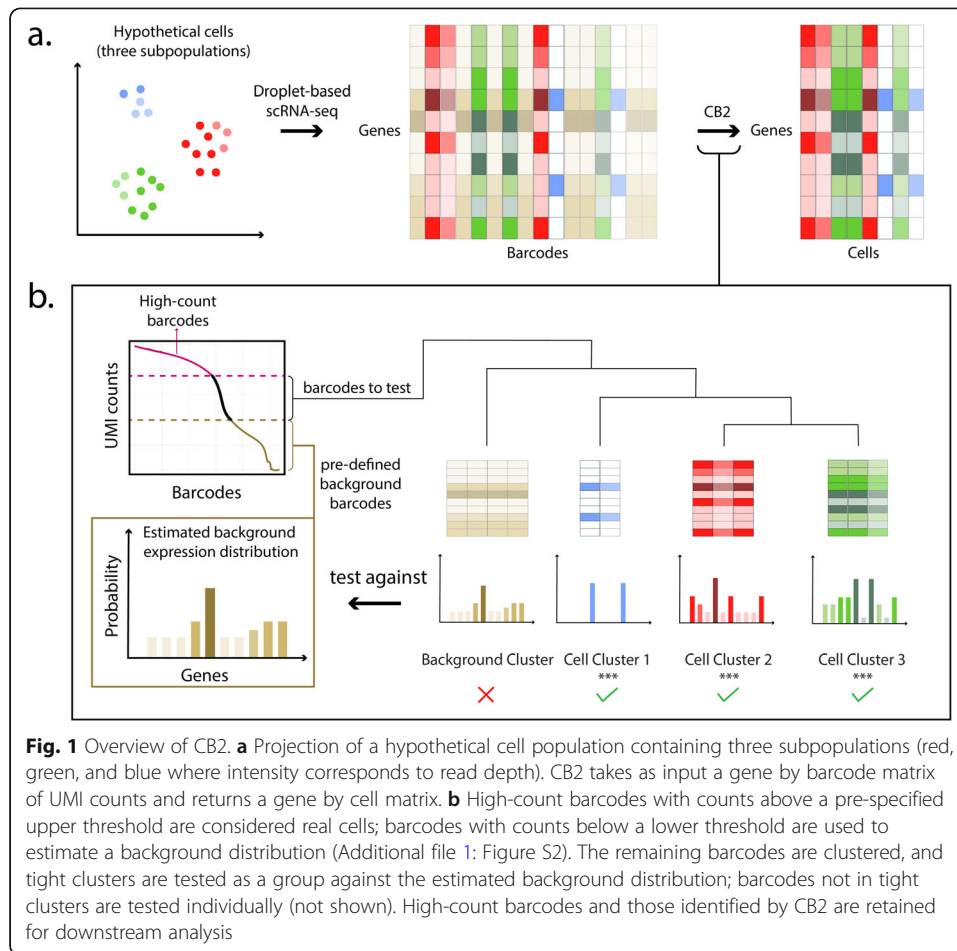
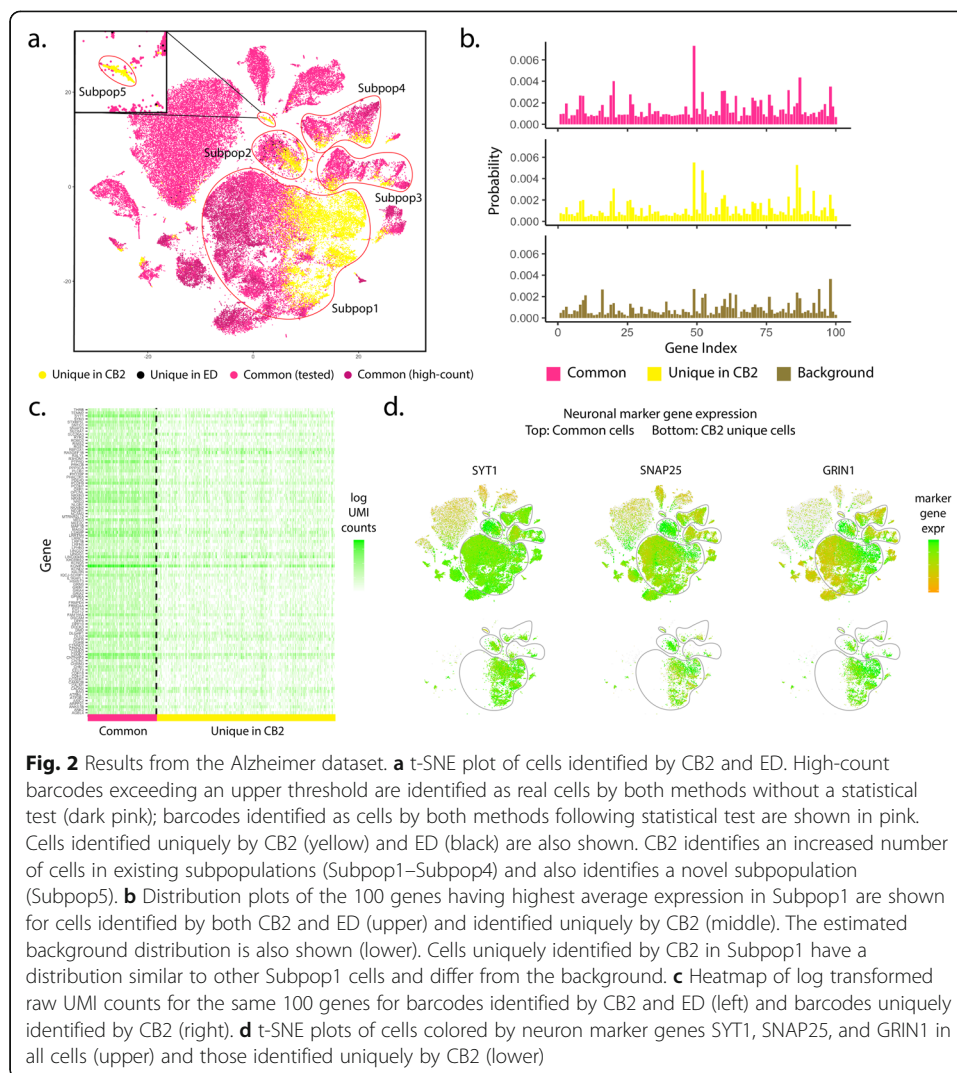


Fig. 1 Overview of CB2. **a** Projection of a hypothetical cell population containing three subpopulations (red, green, and blue where intensity corresponds to read depth). CB2 takes as input a gene by barcode matrix of UMI counts and returns a gene by cell matrix. **b** High-count barcodes with counts above a pre-specified upper threshold are considered real cells; barcodes with counts below a lower threshold are used to estimate a background distribution (Additional file 1: Figure S2). The remaining barcodes are clustered, and tight clusters are tested as a group against the estimated background distribution; barcodes not in tight clusters are tested individually (not shown). High-count barcodes and those identified by CB2 are retained for downstream analysis

(range 44–100%) add to existing subpopulations. The remaining 12% (range 0–56%) make up novel subpopulations.

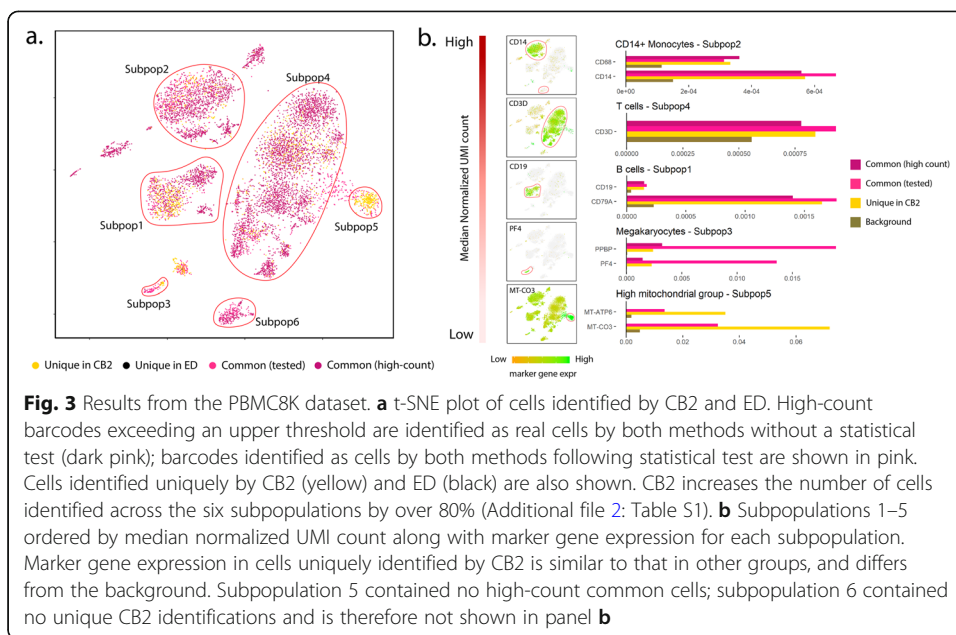
As an example, Fig. 2 and Additional file 1: Figure S6 show results from the Alzheimer data [5] where CB2 identifies 18% more cells. A detailed look at the unique CB2 identifications suggests that the extra cells identified are not false positives, but rather they add to existing excitatory neuron and inhibitory neuron subpopulations, and also reveal a novel subpopulation consisting of 209 cells. Specifically, Fig. 2 b and c show distribution plots and an expression heatmap of the 100 genes having the highest average expression in Subpop1 (the largest subpopulation) for cells identified by both CB2 and ED as well as those identified uniquely by CB2. As shown, cells uniquely identified by CB2 have a distribution similar to other cells, and they differ from the background. Using the marker genes from Mathys et al. [5], Fig. 2d and Additional file 1: Figure S6(b) suggest that cells identified uniquely by CB2 in Subpops 1–4 are neurons, as they show relatively high expression of neuron marker genes SYT1, SNAP25, and GRIN1. More specifically, the CB2 cells in Subpops 1–2 exhibit high expression of excitatory neuronal markers whereas the cells in Subpops 3–4 appear to be inhibitory neurons (Additional file 1: Figure S6(c) and (d)). The novel subpopulation (Subpop5) uniquely shows high expression of both oligodendrocyte and astrocyte marker genes, suggesting



that this group may be mixed phenotype glial cells [6] (Additional file 1: Figure S6(e) and (f)).

By increasing the number of real cells identified, CB2 also improves the power to differentiate Alzheimer's patients from controls. Specifically, Mathys et al. [5] profiled expression from the prefrontal cortex of 24 AD-pathology patients as well as 24 age-matched controls, and they validated differentially expressed genes in different cell types, including 9 genes in excitatory neurons and 9 in inhibitory neurons. Additional file 1: Figure S7 shows that by identifying additional cells, CB2 improves downstream differential expression analysis by resulting in more significant p values and stronger fold changes.

In a second case study (PBM8K), CB2 increases the number of cells identified across six subpopulations by over 80% (Additional file 2: Table S1). Results are shown in Fig. 3 and Additional file 1: Figure S8. Similar to the Alzheimer's data analysis, Additional file 1: Figure S8(b) and (c) show that cells identified uniquely by CB2 in Subpop1 have an expression profile that is similar to other cells and differs from the background. Figure 3 provides a detailed look at marker gene expression for the



well-characterized PBMC8K cells using markers considered in Zheng et al. [2]. As shown in Fig. 3b, CB2 identifies additional CD14+ monocytes, T cells, B cells, and megakaryocytes. Results from two additional datasets are shown in Additional file 1: Figure S9–S10.

Discussion

Taken together, the results presented here demonstrate that CB2 provides a powerful approach for distinguishing real cells from background barcodes which will increase the number of cells identified in existing cell subpopulations in most datasets and may facilitate the identification of novel subpopulations. While advantages are expected in many settings, users will benefit from the following considerations. CB2 does not test for doublets or multiplets, and consequently, some of the high-count identifications may consist of two or more cells. Methods for identifying multiplets such as Scrublet [7], DoubletDecon [8], or DoubletFinder [9] may prove useful after applying CB2. A second important post-processing step is filtering based on mitochondrial expression. As noted in Lun et al. [4], any method for distinguishing cells from background barcodes is technically correct in identifying low-quality cells given that damaged cells exhibit expression profiles that differ from the background. Specifically, mitochondrial gene expression is often high in damaged cells; an example is shown in Subpopulation 5 of the PBMC8K data (Fig. 3b). Such cells are typically not of interest in downstream analysis and should therefore be removed. The *GetCellMat* function in *R/scCB2* may be used toward this end.

Conclusions

Droplet-based scRNA-seq technologies provide unprecedented opportunity to address biological questions, but efficient pre-processing is required to maximize the information obtained in an experiment. CB2 allows investigators to maximize the number of cells retained and consequently to increase the power and precision of downstream analysis.

Methods

Versions

For cell identification with R/*scCB2* 0.99.12 and R/*DropletUtils* 1.5.4 [4, 10], the latest version of R [11] was used: 3.7-devel (2019-07-17 r76847). Other packages are not yet compatible or not stable with the R developers version, and so for *scrna* 1.12.1 [12], *Seurat* 3.1.0 [13, 14], and *ggplot2* 3.2.1 [15], R 3.6.0 (2019-04-24 r76423) was used.

CB2

As CB2 relies on ED, we briefly review the ED approach before detailing the clustering test introduced in CB2. ED expects as input a $G \times B$ feature-by-barcode matrix with G features (for simplicity, we refer to features as genes) and B barcodes. Barcodes having zero counts for all genes are filtered out, and the remaining barcodes are divided into three groups based on the sum of gene expression (UMI) counts within a barcode. The background group, B_0 , contains all barcodes with counts less than or equal to a pre-defined lower threshold (defaults to 100); the high-count barcodes, B_2 , contain barcodes with counts exceeding an upper threshold (defaults to knee point); the remaining barcodes (B_1) are tested (Additional file 1: Figure S2).

ED assumes that counts from a background barcode are distributed as Dirichlet-Multinomial with probability vector p_{B_0} estimated by averaging the counts in B_0 and applying the Good-Turing algorithm [16] to ensure that all probabilities are non-zero, denoted as \hat{p}_{B_0} . For a barcode $b \in B_1$, ED tests $p_b = p_{B_0}$ against the alternative $p_b \neq p_{B_0}$ using the log-likelihood under \hat{p}_{B_0} as the test statistic. A Monte-Carlo p value is calculated via simulating Dirichlet-Multinomial barcodes of size $|b|$ under \hat{p}_{B_0} and calculating the proportion of simulated barcodes having a test statistic more extreme than (or equal to) b 's. The false discovery rate is controlled using the Benjamini-Hochberg procedure [17].

CB2 follows ED by filtering out genes with zero counts and dividing the remaining barcodes into three groups. However, instead of testing all barcodes from B_1 individually, CB2 first clusters barcodes and then tests tight clusters to identify those that differ from the background. As in methods for genome-wide association studies (Mieth et al. 2016 [18]), gene co-expression network analysis (Botía et al., 2017 [19]), and de novo transcriptome analysis (Malik et al., 2018 [20]), clustering prior to testing increases power by reducing the total number of tests and increasing the signal to noise ratio. CB2 proceeds as follows:

1. Barcodes grouped by size. CB2 orders barcodes in B_1 by total counts

$$B_1 = \{b_1, \dots, b_{|B_1|}\} \text{ s.t. } |X_{b_i}| \leq |X_{b_{i+1}}|$$

where X_b denotes the count vector of barcode b , $|X_b|$ denotes the total UMI count of barcode b , and $|B_1|$ denotes the number of barcodes in B_1 . Groups of size S (defaults to 1000 in R/*scCB2*) are constructed consisting of barcodes ranging in size from smallest to largest:

$$B_{11} = \{b_1, \dots, b_S\}, B_{12} = \{b_{S+1}, \dots, b_{2S}\}, \dots, B_{1K} = \{b_{(K-1)S+1}, \dots, b_{|B_1|}\}$$

where $K = \frac{|B_1|}{S}$ is rounded up if not an integer. If $|B_{1K}| < \frac{S}{2}$, barcodes in B_{1K} are merged with those in $B_{1(K-1)}$. Sorting barcodes by size reduces bias in the clustering and testing steps that follow.

2. Barcodes clustered within group: Barcodes within each group B_{1j} are clustered using hierarchical clustering with pairwise Pearson correlation as the similarity metric. A cluster is considered tight if the average within-cluster pairwise Pearson correlation exceeds a data-driven threshold. Tight clusters are retained for further analysis as described in step 3, below. To determine thresholds, ten tight clusters of varying size are simulated by generating 100 samples from a multinomial distribution with parameters (N, p) where N ranges from 100 to 1000 in increments of size 100. This range is chosen as we found little variation in thresholds for barcode sizes exceeding 1000; p is set to either \hat{p}_{B_0} or \hat{p}_{B_2} , whichever has larger Shannon entropy [21] as the distribution with larger entropy is less affected by outliers. For each simulated cluster C , the threshold κ_C is defined by its average pairwise Pearson correlation. A cluster is considered tight if the average within-cluster pairwise Pearson correlation exceeds κ_C for the simulated cluster of closest size.
3. Tight clusters tested: For each tight cluster C , we conduct a Monte-Carlo test to assess dissimilarity from the background. Pairwise Pearson correlations are calculated between every barcode in C and \hat{p}_{B_0} ; the test statistic for cluster C , T_C , is defined to be the median of these correlations. Similar to ED, to simulate background barcodes, we sample barcodes X_1^*, \dots, X_M^* from a multinomial $(N; \hat{p}_{B_0})$ where N is the size of the barcode giving T_C . The Monte-Carlo p value is:

$$p_C = \frac{\sum_{i=1}^M \{cor_{X_i^*,0} \leq T_C\} + 1}{M + 1}$$

where $cor_{X_i^*,0}$ is the Pearson correlation between X_i^* and \hat{p}_{B_0} (M defaults to 1000 in R/scCB2). Monte-Carlo p values are calculated for each cluster followed by Benjamini-Hochberg [17] to control the FDR. All barcodes within a significant cluster are identified as real cells.

4. Individual barcodes tested: Barcodes that were not included in a tight cluster in Step 2 as well as those in a tight cluster that were not found to be significant in Step 3 are tested individually using ED. It is important to note that some of the barcodes identified in this step do not overlap with identifications made when ED is applied to the full set of barcodes given differences in the rates of real cells to background barcodes and differences in error rate control.

Simulations

Counts are generated as in Lun et al. [4]. As detailed there, each simulation requires an input dataset. We constructed simulations from 10 datasets: Alzheimer [5], PBMC8K, PBMC33K, mbrain1K, mbrain9K, PanT4K, MALT, PBMC4K, jurkat, and T293 (Additional file 2: Table S2). For each input dataset, the inflection point of the UMI count by sorted barcode plot is used to divide lower count from higher count barcodes. The barcodes in the lower count range are considered background. In SIM IA, two sets of

2000 barcodes randomly sampled from the higher count range are considered real cells. The first set of 2000 is referred to as large (G_1) cells; the second set is downsampled by 90% to give small (G_2) cells. We added a third set of medium ($G_{1.5}$) cells by sampling 2000 cells from the higher count range and downsampling by 50%. The process for simulating data in SIM IB is identical to SIM IA except that in SIM IB, 10% of the genes in each simulated real cell are shuffled making the real cells more different from the background barcodes and, consequently, making real cells easier to identify. SIM IA is a more realistic simulation (Additional file 1: Figure S4).

Case studies

We evaluated the 10 datasets used in the simulation and also the placenta data evaluated in Lun et al. [4]. These datasets vary in sequencing depth as well as in the extent of differences between the real cell and background distributions (Additional file 1: Figure S4). CB2 and ED were applied to each dataset using default settings. For plots that compare identifications between CB2 and ED, cells identified by either approach (or both) were combined and UMI counts were normalized via *scran*. The *Seurat* pipeline was used to generate t-SNE plots from the top 4000 most highly variable genes and top 50 principal components. Expression heatmaps show log transformed raw UMI counts. For heatmaps and distribution plots, mitochondrial and ribosomal genes were removed.

Differential expression analysis in Alzheimer data

Cells identified by CB2, ED, or both were combined into a single matrix and filtered similar to Mathys et al. [5]. Specifically, cells with mitochondrial gene expression making up 40% or more of the total UMI counts were removed; genes detected in fewer than two cells were also excluded giving a matrix of 28,208 genes and 74,579 barcodes. Normalization was performed using *scran*. Cell types were annotated using marker genes as in Mathys et al. [5]. Differential expression (DE) tests between cells from Alzheimer's cases and controls were conducted using Wilcoxon rank-sum tests as in Mathys et al. [5]. Results were compared for known DE genes extracted from Mathys et al. [5].

Implementation of CB2 and ED

For all simulation and case study analyses, CB2 and ED were implemented using default parameters. A target FDR was set at 1%.

Existing subpopulations vs. novel subpopulations

The *FindNeighbors* and *FindClusters* functions in *Seurat* were used with default settings to assign each cell to a cluster, referred to here as a subpopulation. For each subpopulation, we calculated the percentage of cells identified by both CB2 and ED as well as those identified uniquely by CB2. Subpopulations for which over 80% of the cells are uniquely identified by CB2 are referred to as novel subpopulations (Additional file 2: Table S3 shows the number of novel subpopulations identified using 70%, 80%, or 90% as thresholds).

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02054-8>.

Additional file 1: Figure S1. Overview of droplet-based single-cell RNA-seq protocol. **Figure S2.** Visualization of barcode groups in the UMI counts v.s. barcodes plot. **Figure S3.** Simulation results under setting SIM IA. **Figure S4.** Comparison between two simulation settings. **Figure S5.** Simulation results under setting SIM IB. **Figure S6.** Additional analysis of Alzheimer dataset. **Figure S7.** Differential Expression analysis between Alzheimer's disease (AD) cases and controls. **Figure S8.** Analysis of PBMC8K dataset. **Figure S9.** Analysis of mbrain1K dataset. **Figure S10.** Analysis of placenta dataset.

Additional file 2: Table S1. The number of cells identified by CB2, ED, or both in 11 case study datasets. **Table S2.** Links to all datasets used in this study. **Table S3.** Number of novel subpopulations identified by CB2 in each dataset.

Additional file 3. Review history.

Acknowledgements

The authors thank Matt Bernstein, Chitrasen Mohanty, and Ziyue Wang for comments that improved the manuscript.

Review history

The review history is available as Additional file 3.

Peer review information

Barbara Cheifet was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Z.N. and C.K. designed the research, developed the method, and wrote the first version of the manuscript. Z.N. and S.C. conducted simulations and quality control evaluations. Z.N., S.C., and C.K. built and tested the R package. Z.N., J.B., and C.K. analyzed results from early versions of the method which helped during method refinement. All authors contributed to writing the manuscript. The authors read and approved the final manuscript.

Funding

This work was supported by NIH GM102756.

Availability of data and materials

The Alzheimer case study dataset was downloaded from <https://www.synapse.org/#!Synapse:syn16780177> [5]. The placenta dataset [22] is available at <https://jmlab-gitlab.cruk.cam.ac.uk/publications/EmptyDrops2017-DataFiles>. All other datasets in this study are available at the 10x Genomics website (<https://support.10xgenomics.com/single-cell-gene-expression/datasets>) (Additional file 2: Table S2). The R package R/scCB2 is available at <https://github.com/zijianni/scCB2> [23] under the General Public License version 3, and will be submitted to Bioconductor [24]. All simulation codes and case study data analysis scripts are available at <https://github.com/zijianni/codes-for-CB2-paper> and Zenodo (doi: 10.5281/zenodo.3829649) [25].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA. ²Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA.

Received: 10 December 2019 Accepted: 21 May 2020

Published online: 08 June 2020

References

1. Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161:1202–14.
2. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
3. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*. 2014;11:163–6.
4. 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*. 2019;20:63.
5. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*. 2019;570:332–7.

6. Dyer CA, Kendler A, Jean-Guillaume D, Awatramani R, Lee A, Mason LM, et al. GFAP-positive and myelin marker-positive glia in normal and pathologic environments. *J Neurosci Res*. 2000;60:412–26.
7. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*. 2019;8:281–291.e9.
8. DePasquale EAK, Schnell DJ, Van Camp PJ, Valiente-Alandí Í, Blaxall BC, Grimes HL, et al. DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. *Cell Rep*. 2019;29:1718–1727.e8.
9. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell Syst*. 2019;8:329–337.e4.
10. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat Commun*. 2018;9:2667.
11. R Core Team. R: A Language and Environment for Statistical Computing 2019.
12. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*. 2016;5:2122.
13. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36:411–20.
14. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177:1888–1902.e21.
15. Wickham H. ggplot2: Elegant Graphics for Data Analysis; 2016.
16. Gale WA, Sampson G. Good-turing frequency estimation without tears*. *J Quant Linguist*. 1995;2:217–37.
17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57:289–300.
18. Mieth B, Kloft M, Rodríguez JA, Sonnenburg S, Vobruba R, Morcillo-Suárez C, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci Rep*. 2016;6:1–14.
19. Botía JA, Vandrovцова J, Forabosco P, Guelfi S, D'Sa K, Hardy J, et al. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Syst Biol*. 2017;11:47.
20. Malik L, Almodaresi F, Patro R. Grouper: graph-based clustering and annotation for improved de novo transcriptome analysis. *Bioinformatics*. 2018;34:3265–72.
21. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–423.
22. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*. 2018;563:347–53.
23. Ni Z, Chen S, Brown J, Kendziorski C. CB2 improves power of cell detection in droplet-based single-cell RNA sequencing data. GitHub. 2020. <https://github.com/zijanni/scCB2>. Accessed 15 May 2020.
24. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*. 2015;12:115–21.
25. Ni Z, Chen S, Brown J, Kendziorski C. CB2 improves power of cell detection in droplet-based single-cell RNA sequencing data. Zenodo. 2020; <https://doi.org/10.5281/zenodo.3829649>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

