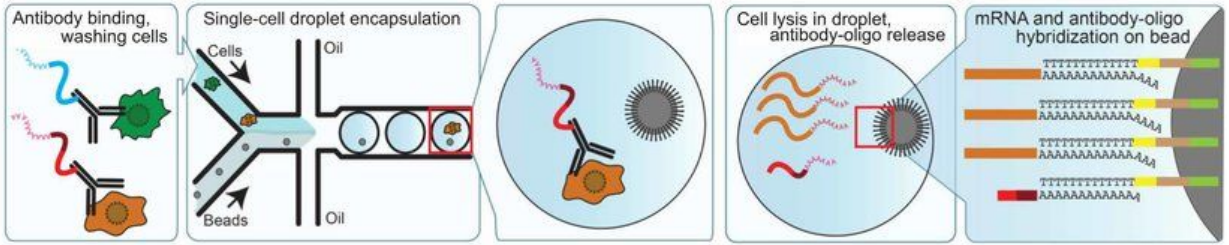
INJECTING RIGOR AND REPRODUCIBILITY INTO *CITE-SEQ* WORKFLOWS:
DECONTAMINATION AND IN SILICO GATING APPROACHES

Jamie Park/Ava Jensen, Tim Triche
Van Andel Institute Graduate School

CITE-seq = scRNAseq + surface protein expression

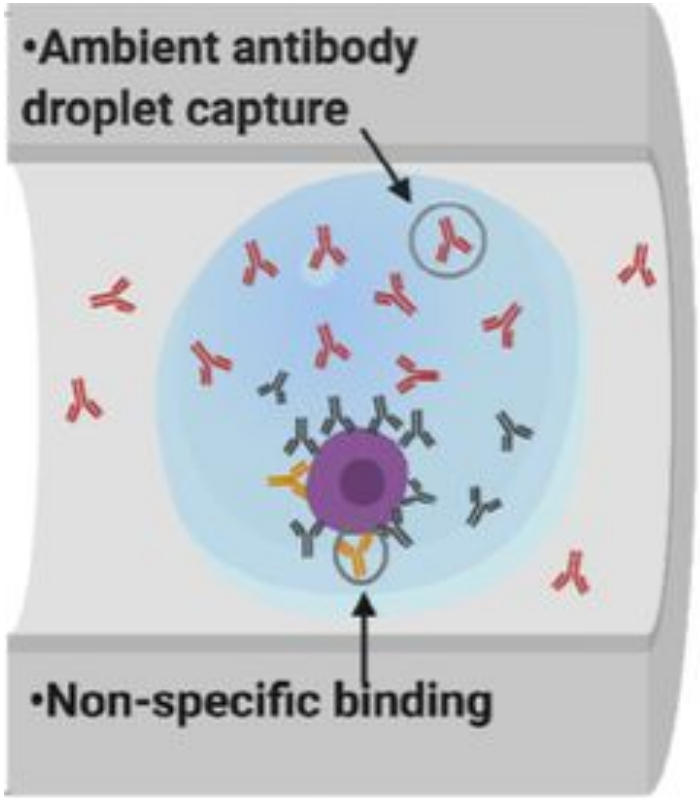
Cellular **I**ndexing of **T**ranscriptomes and **E**pitopes.

ADT (*antibody derived tags*) allow co-sequencing of transcriptome + cell surface proteins.



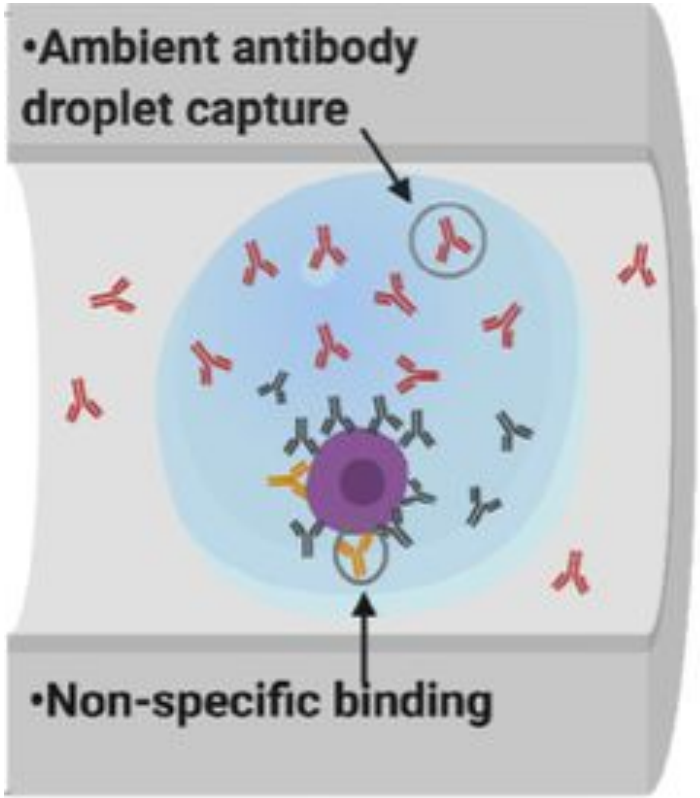
Stoeckius, et al., 2017

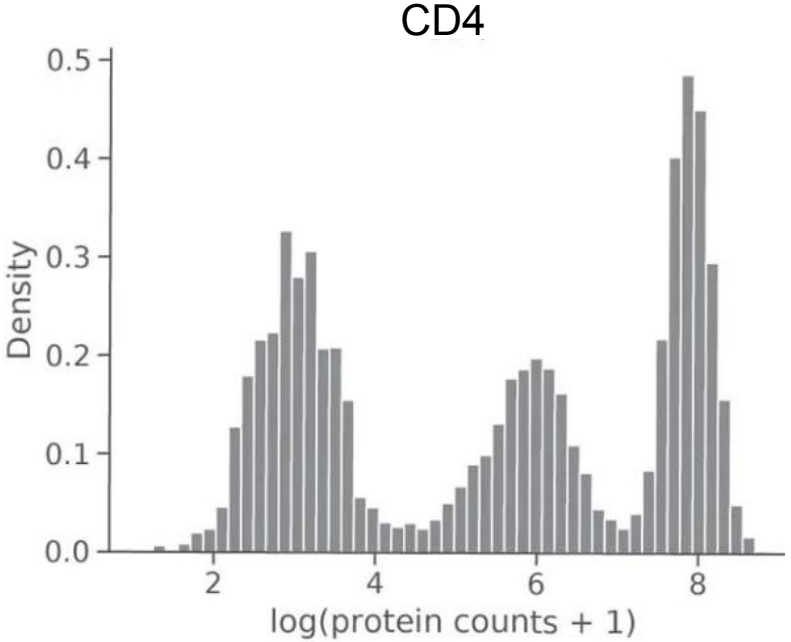
CITE-seq data is noisy.



Gayoso et al., 2021

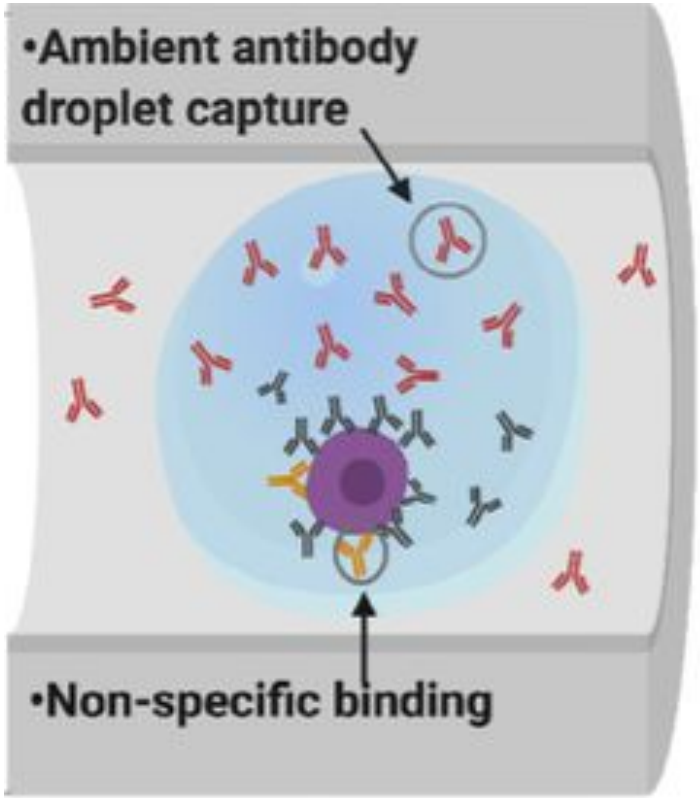
CITE-seq data is noisy.

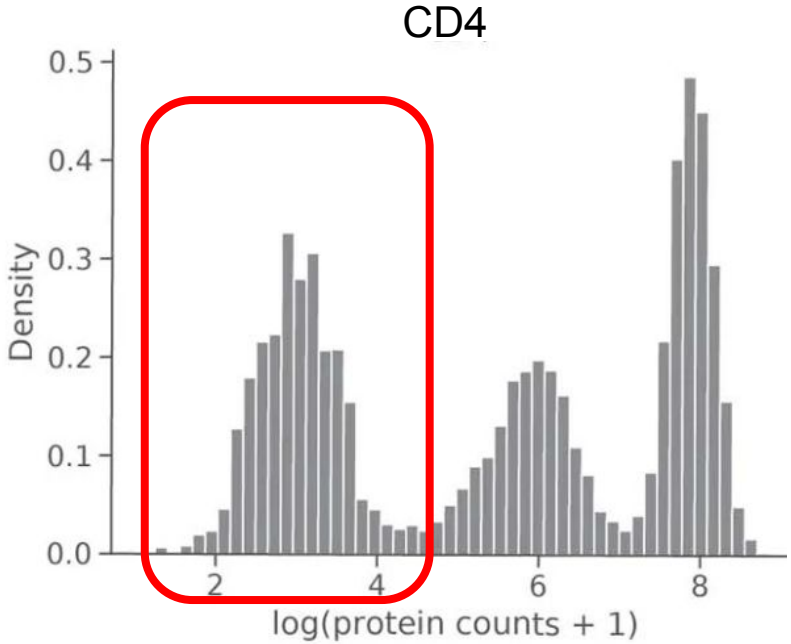




Gayoso et al., 2021

CITE-seq data is noisy.





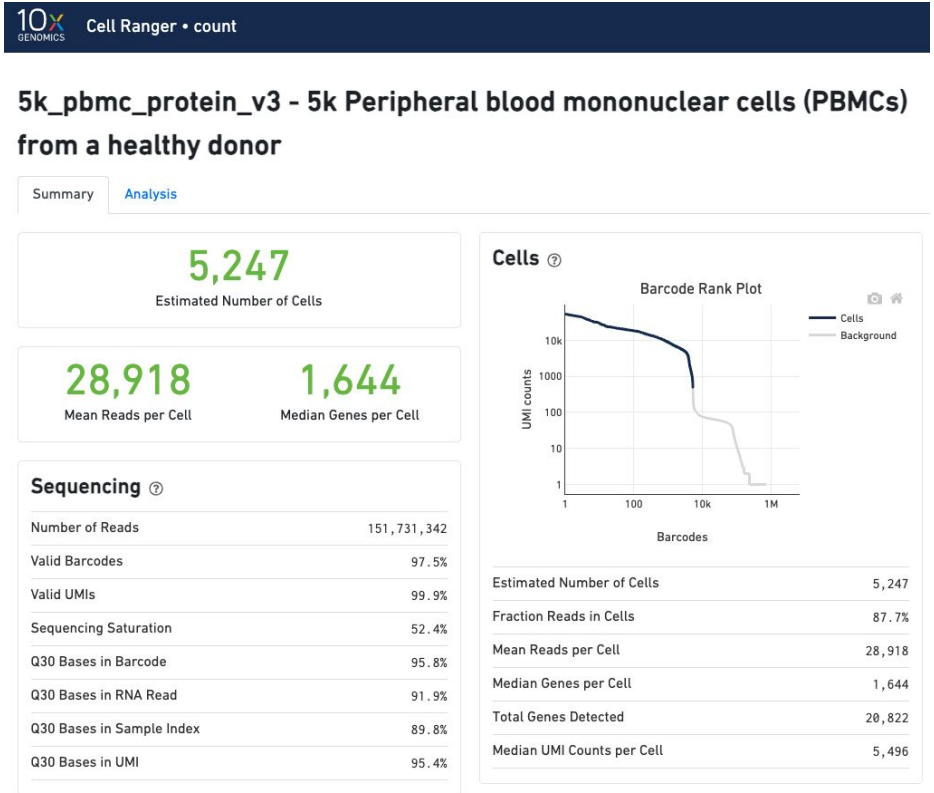
CD4 expression of PBMC CITE-seq contain 3 modes (Background, Monocytes, and CD4+ T cells)

Gayoso et al., 2021

Packages for decontamination in CITE-seq assays

pkg name	language	Pubs
dsb	**R**	**Mulè, et al., 2022**
decontX	**R**	**Yang et al., 2020**
totalVI	python	Gayoso et al., 2021
scAR	python	Sheng et al., 2022 (*bioRxiv*)

Using PBMC_5K on Macbook (M1) for reference



- 5k_pbmc_protein (v3) is a publicly available dataset from Genomics 10X

- All packages were run locally on Macbook with 16Gb of RAM.

dsb uses empty droplet ADT expression to estimate noise.

2 types of noise :

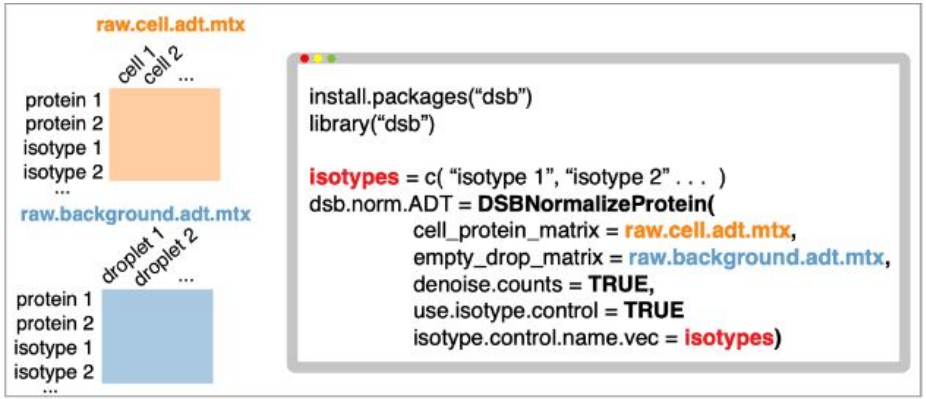
Article | Open Access | Published: 19 April 2022

Normalizing and denoising protein expression data from droplet-based single cell profiling

Matthew P. Mulè, Andrew J. Martins & John S. Tsang

Nature Communications **13**, Article number: 2099 (2022) | Cite this article

12k Accesses | **18** Citations | **32** Altmetric | Metrics



Mule.et al, 2022

dsb uses empty droplet ADT expression to estimate noise.

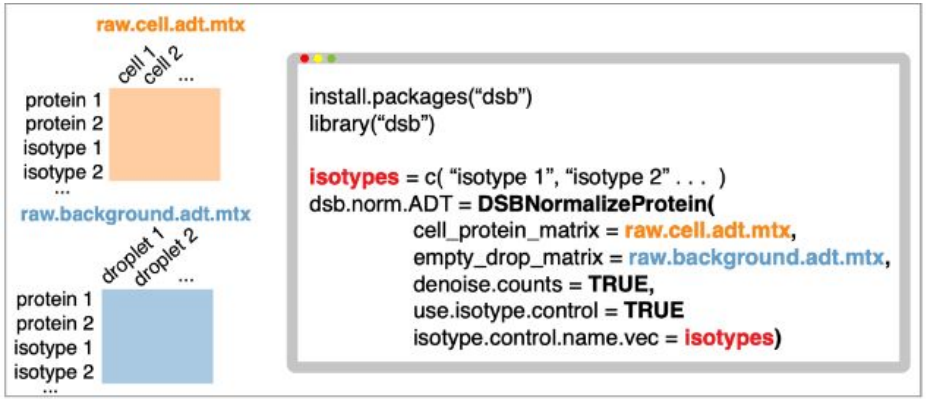
Article | Open Access | Published: 19 April 2022

Normalizing and denoising protein expression data from droplet-based single cell profiling

Matthew P. Mulè, Andrew J. Martins & John S. Tsang

Nature Communications **13**, Article number: 2099 (2022) | Cite this article

12k Accesses | **18** Citations | **32** Altmetric | Metrics



2 types of noise :

1) Protein-specific noise from ambient antibodies

Can be estimated from "emptydrop" raw matrices.

Mule.et al, 2022

dsb uses empty droplet ADT expression to estimate noise.

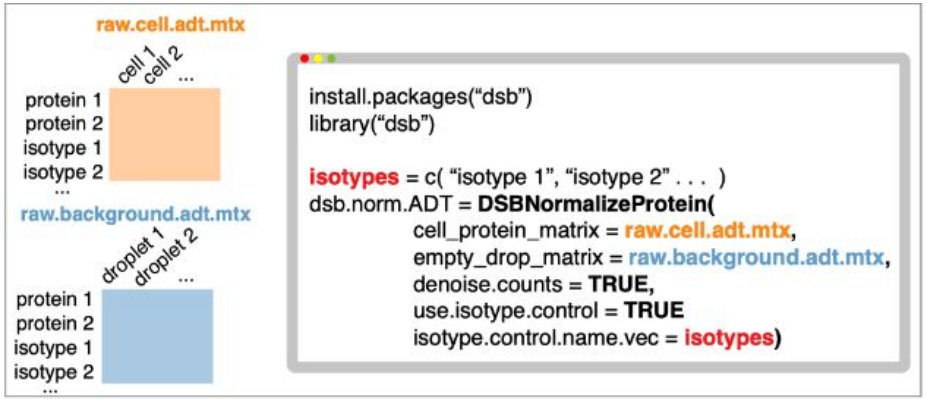
Article | Open Access | Published: 19 April 2022

Normalizing and denoising protein expression data from droplet-based single cell profiling

Matthew P. Mulè, Andrew J. Martins & John S. Tsang

Nature Communications **13**, Article number: 2099 (2022) | Cite this article

12k Accesses | **18** Citations | **32** Altmetric | Metrics



2 types of noise :

1) Protein-specific noise from ambient antibodies

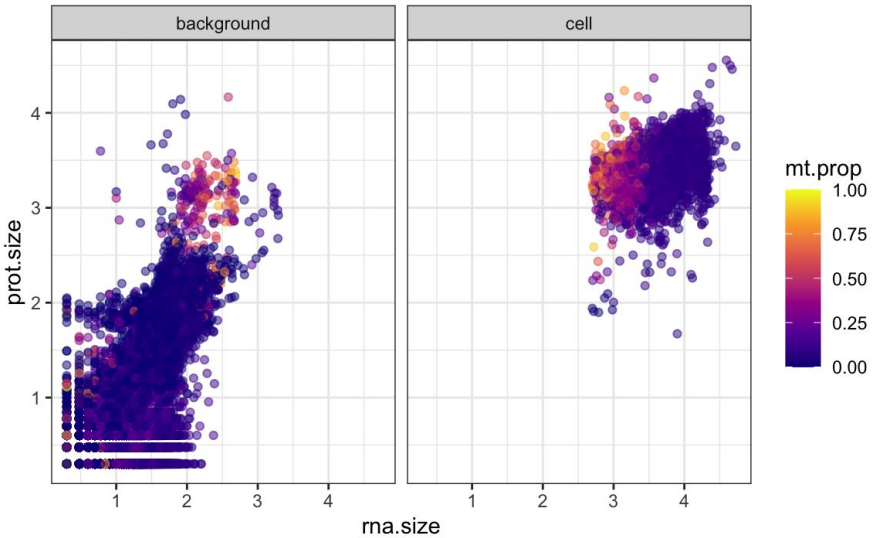
Can be estimated from "emptydrop" raw matrices.

2) droplet/cell-specific noise

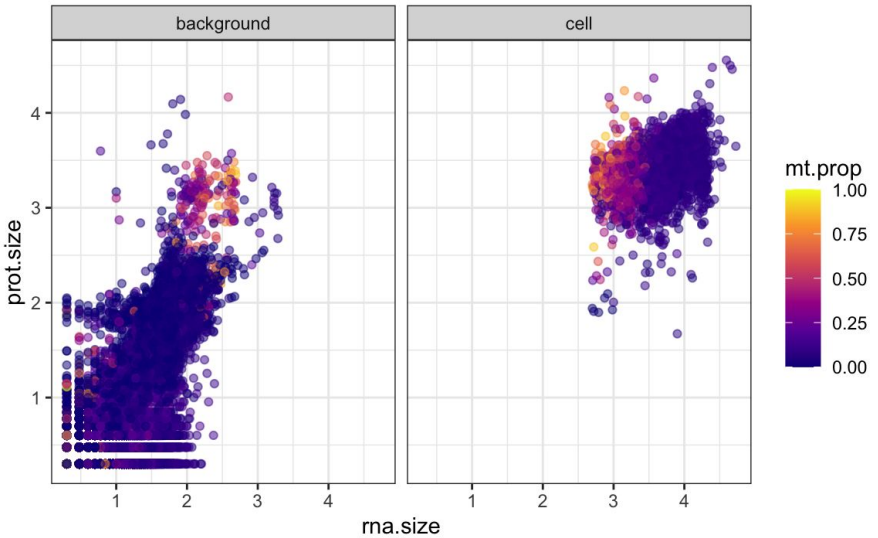
Can be revealed by the shared variance component with isotype antibody controls

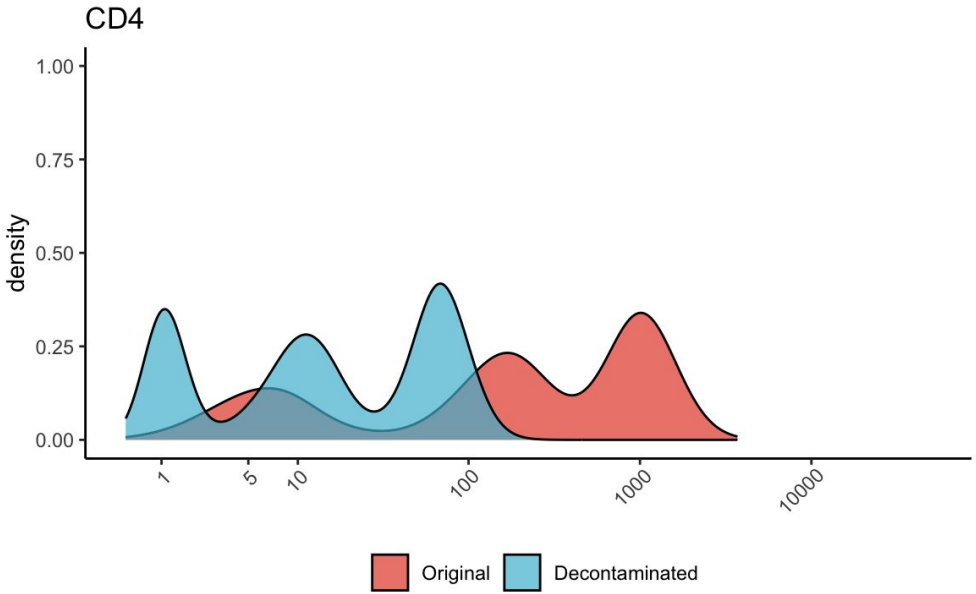
Mule.et al, 2022

dsb decontaminated vs. raw ADT expression (CD4)



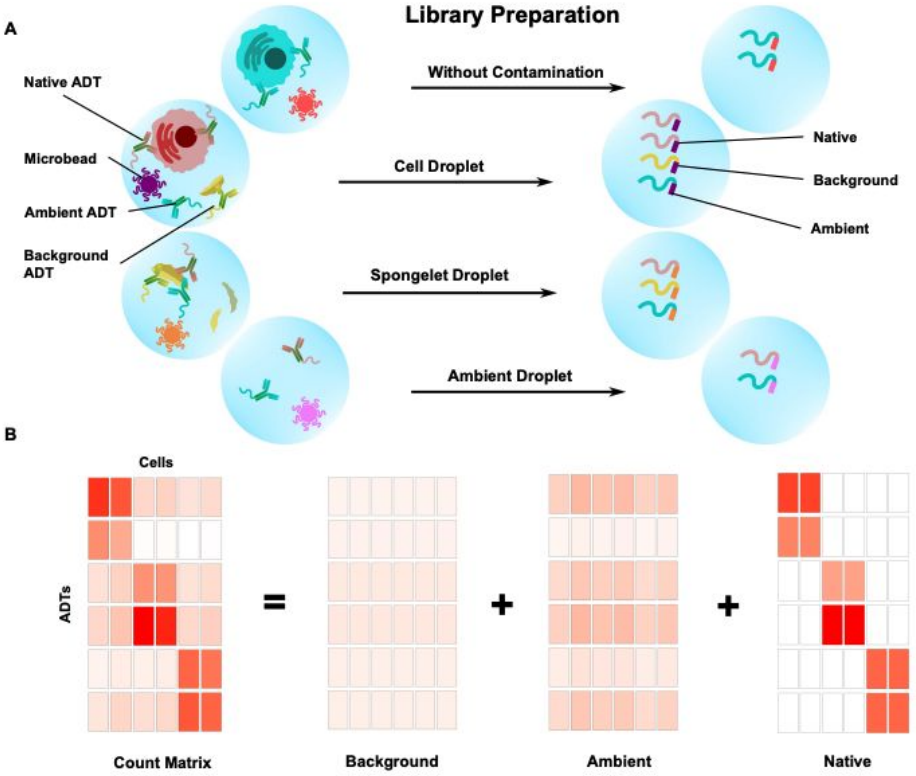
dsb decontaminated vs. raw ADT expression (CD4)





Output matrix was exponentiated for density plot comparison.

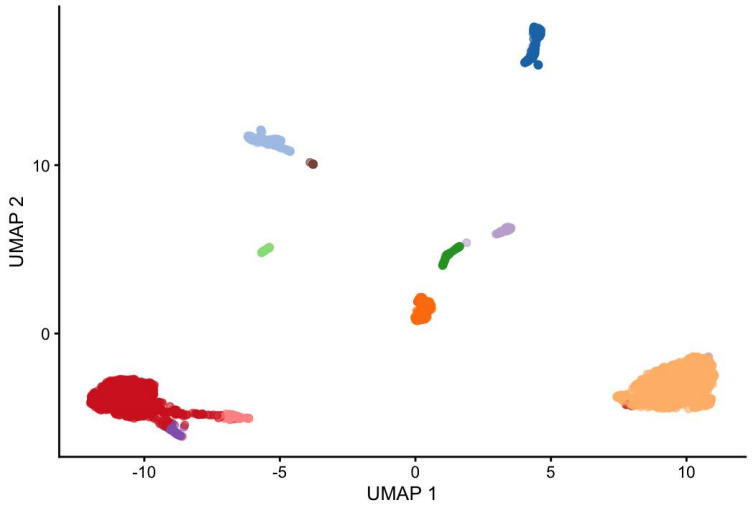
DecontPro also uses spongelets to remove contamination

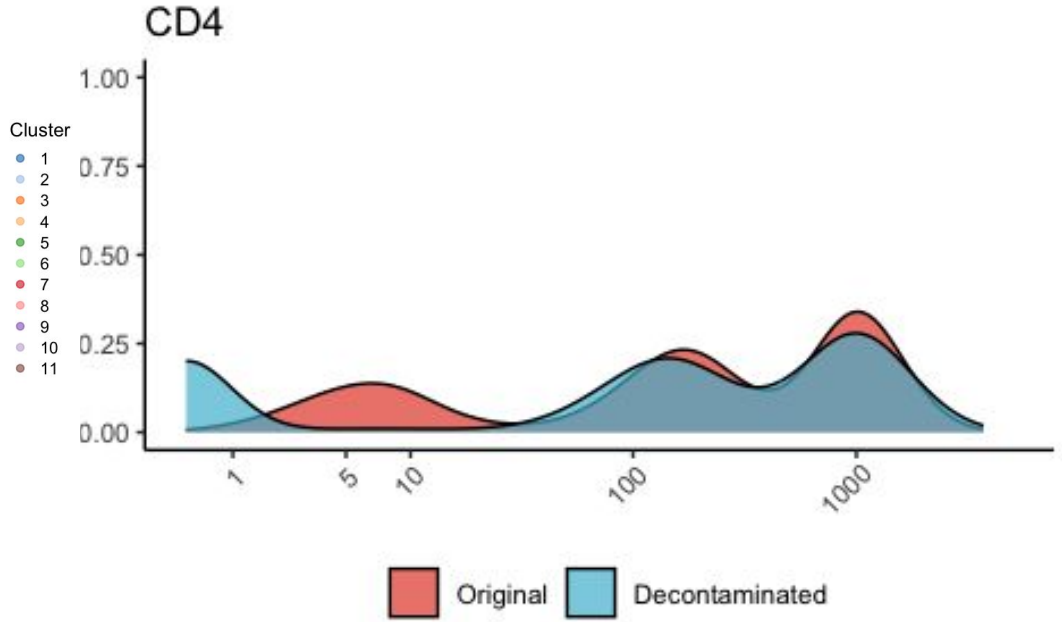


Spongelet : cellular debris with antibody aggregates correlate with expression profiles of the background peak in true cells.

Bayesian hierarchical model to estimate/remove background.

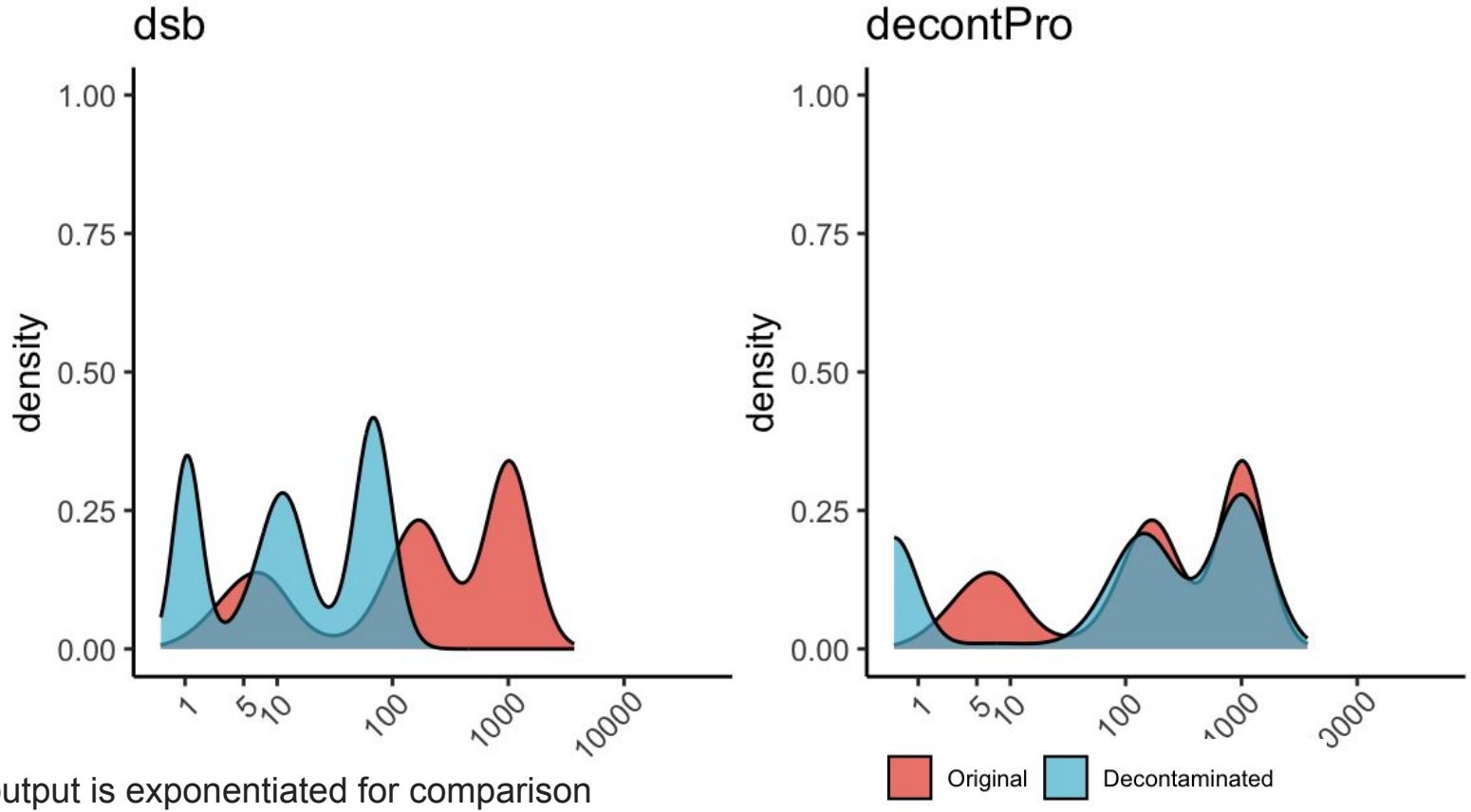
DecontPro decontaminated vs. raw ADT expression (CD4)





decontPro requires cluster info.

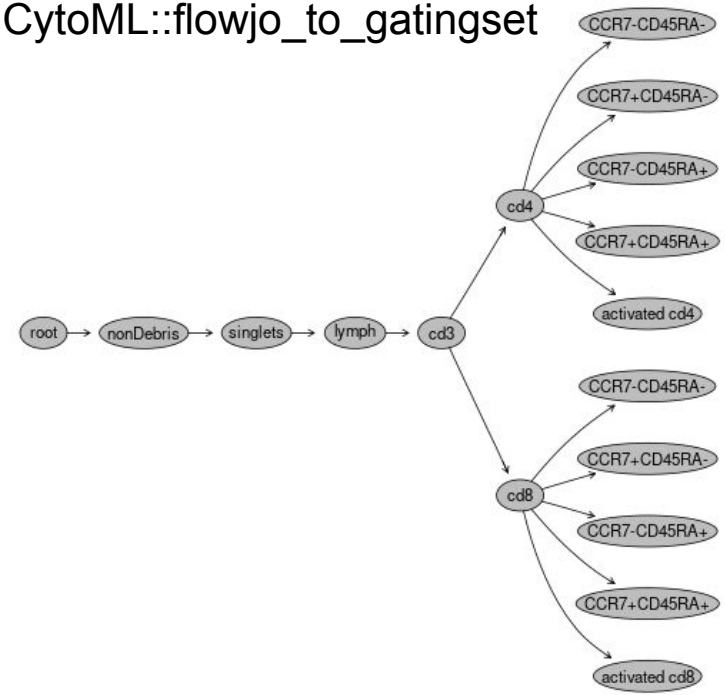
CD4 expression of dsb vs DecontPro decontamination



*dsb output is exponentiated for comparison

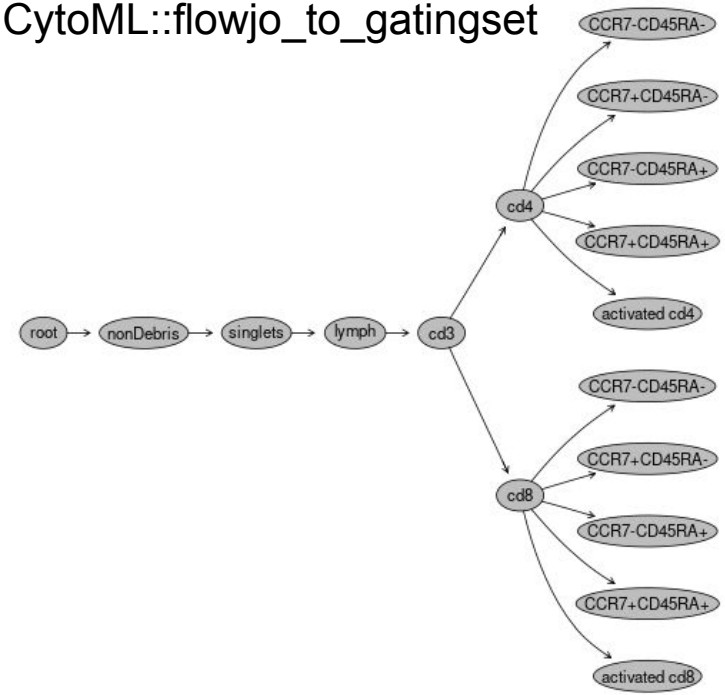
scGate can simulate flow cytometry gating

CytoML::flowjo_to_gatingset

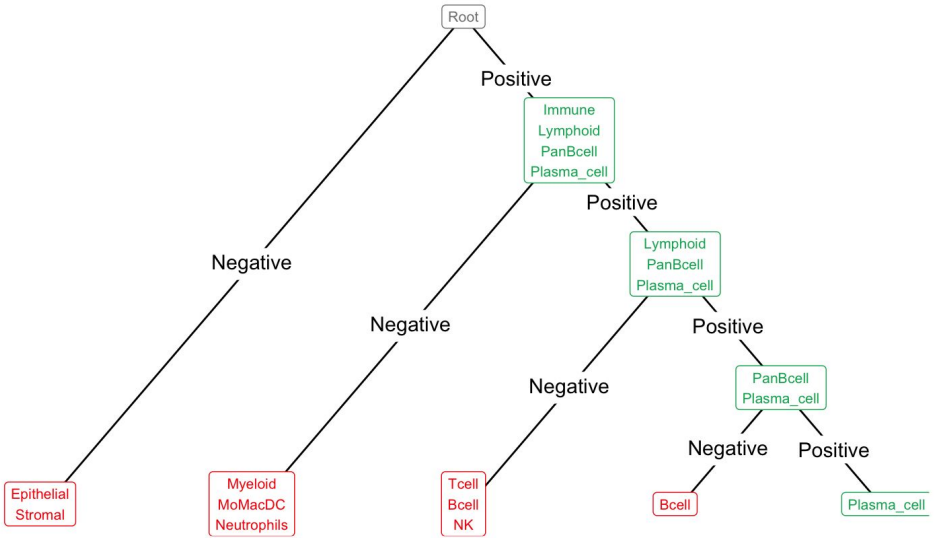


scGate can simulate flow cytometry gating

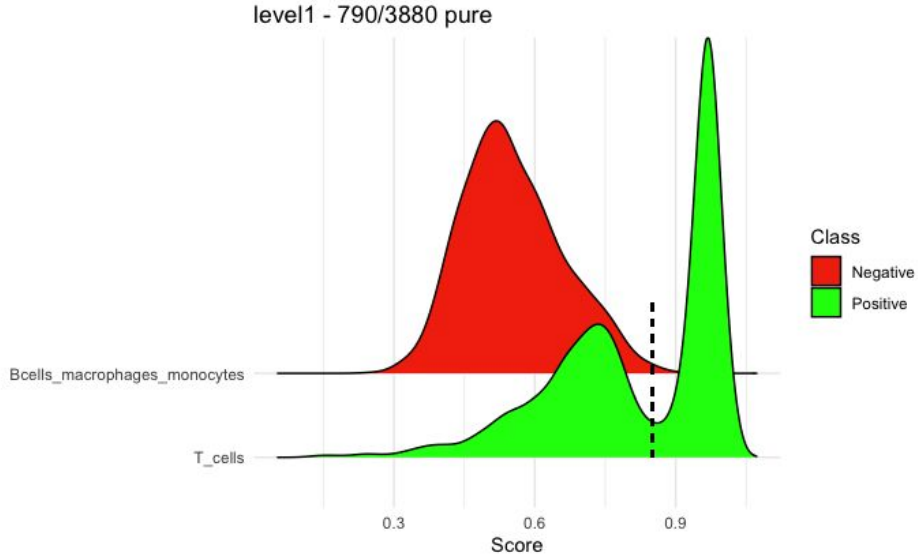
CytoML::flowjo_to_gatingset

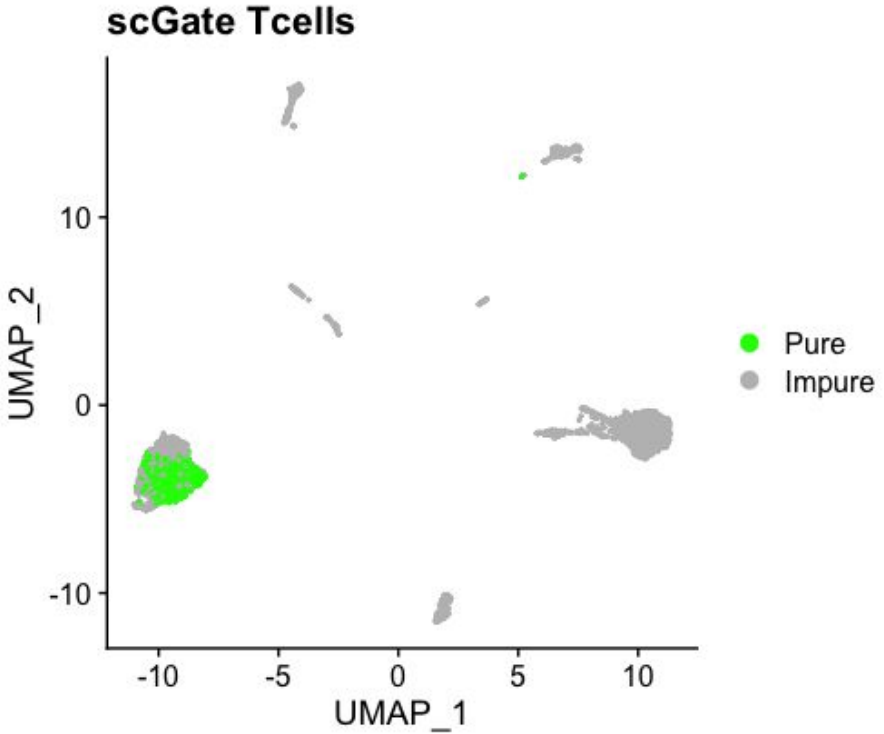


scGate::plot_tree(my_scGate_model)



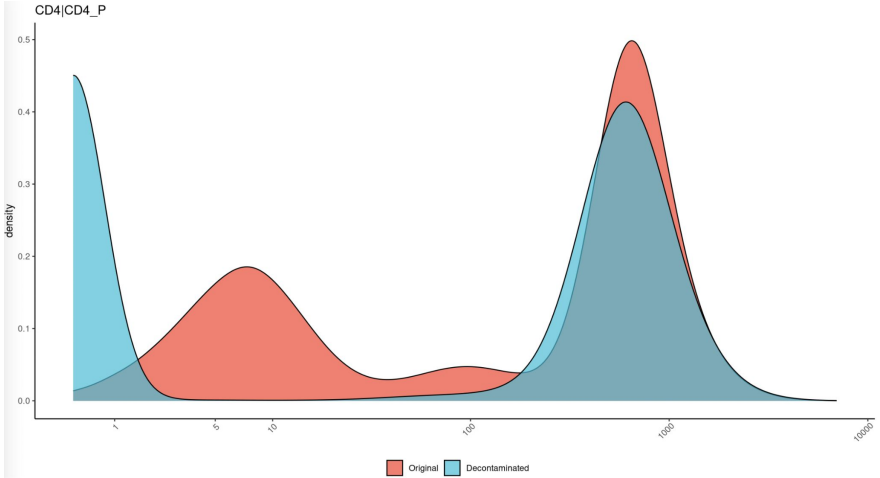
scGate can threshold gates to annotate *pure* T cells

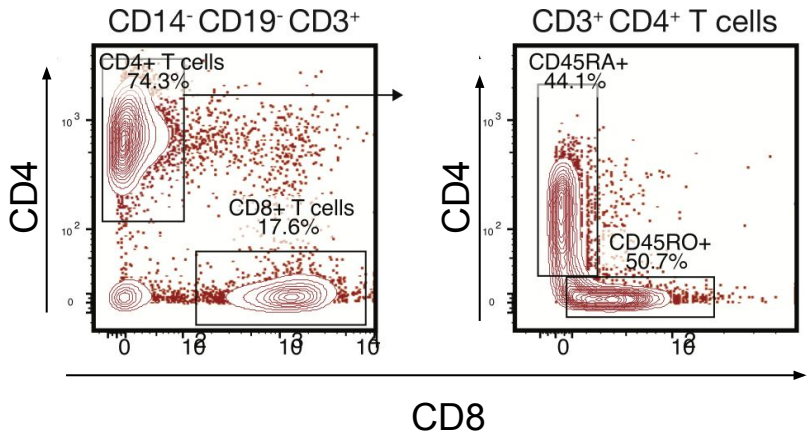




Ucell score based thresholding

Mair et al., 2019 has run Abseq *and* flow in parallel





Conclusions

	decontPro (bioconductor)	Dsb (CRAN)	scGate (CRAN)
Features	- Bayesian hierarchy - Installation challenges - Difficulty to scale without HPC	- Uses emptydroplets - Automates normalization - Performs best with isotype controls	- Automates marker-based purification with in silico gating sets
publication	Yang et al., (bioXriv)	(Mule et al., 2022, Nature Comms)	Andreatta et al., 2022, Bioinformatics)

- Various packages can decontaminate ambient noise in CITE-seq matrices.
  - Dsb uses 1) empty droplets 2) isotype controls to estimate background
  - decontPro uses Bayesian hierarchical algorithms with cell cluster info to estimate contamination.
- scGate can offer *in silico* gating strategies to annotate cells.

Acknowledgements

Tim Triche
Ava Jensen
Lauren Harmon
Zack Ramjan

References

Yang, S., Corbett, S.E., Koga, Y. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol* 21, 57 (2020). https://doi.org/10.1186/s13059-020-1950-6

Gayoso, A., Steier, Z., Lopez, R. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* 18, 272–282 (2021). https://doi.org/10.1038/s41592-020-01050-x

Massimo Andreatta and others, scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets, *Bioinformatics*, Volume 38, Issue 9, March 2022, Pages 2642–2644, https://doi.org/10.1093/bioinformatics/btac141

Stoeckius, M., Hafemeister, C., Stephenson, W. et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods 14, 865–868 (2017). https://doi.org/10.1038/nmeth.4380

Mair F, Erickson JR, Voillet V, Simoni Y, Bi T, Tyznik AJ, Martin J, Gottardo R, Newell EW, Prlic M. A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. Cell Rep. 2020 Apr 7;31(1):107499. doi: 10.1016/j.celrep.2020.03.063. PMID: 32268080; PMCID: PMC7224638.

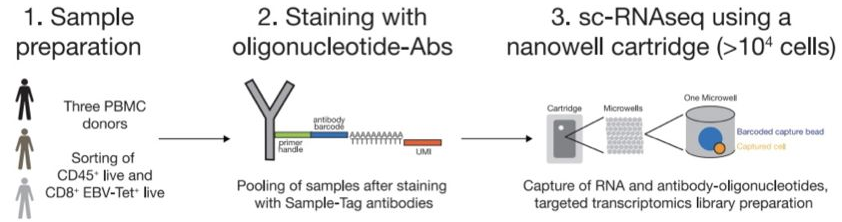
VAN ANDEL INSTITUTE®

Orcid

BACKUPSLIDES

Mostly on LASRY

Mair et al., (Abseq)



Gene expression

scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets

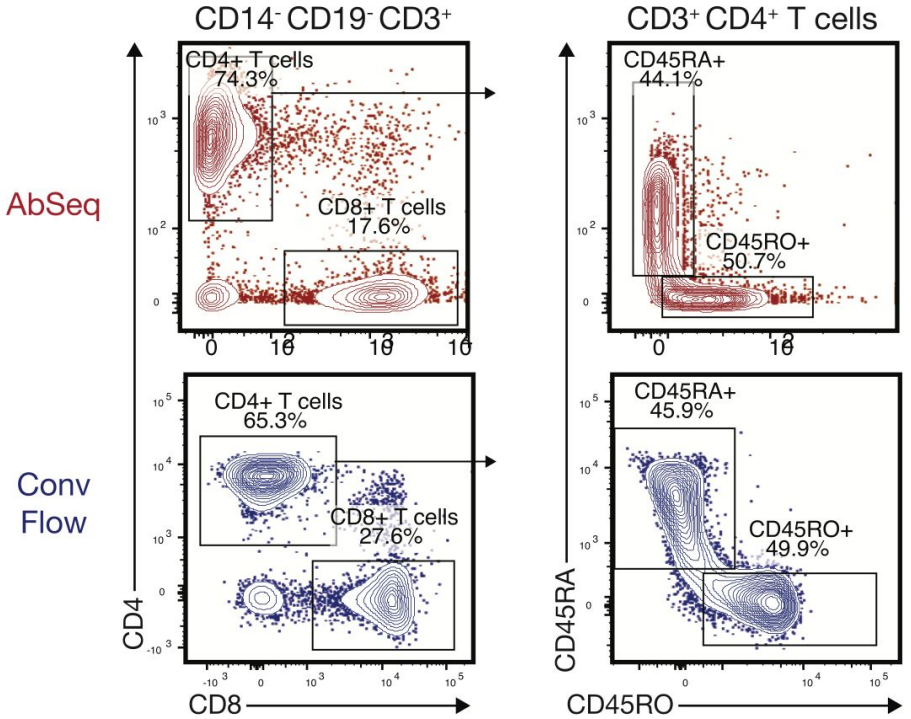
Massimo Andreatta [1,2], **Ariel J. Berenstein** [3] **and Santiago J. Carmona** [1,2,*]

[1]Ludwig Institute for Cancer Research, Lausanne Branch, and Department of Oncology, CHUV and University of Lausanne, 1011 Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland and [3]Laboratorio de Biología Molecular, División Patología, Instituto Multidisciplinario de Investigaciones en Patologías Pediátricas (IMIPP), CONICET-GCBA, Buenos Aires C1425EFD, Argentina

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on November 16, 2021; revised on February 21, 2022; editorial decision on February 28, 2022; accepted on March 4, 2022



Single-cell transcriptomic and proteomic assays have added substantial breadth and depth to our understanding of cellular phenotypes and interactions. Particularly in the study of cellular immunity, the recent CITE-seq and REAP-seq protocols (which simultaneously assay hundreds of cell surface proteins alongside thousands of mRNA transcripts) have provided a robust and scalable means to dissect tissue- and condition-specific roles of individual cells.

However, the most appropriate means to preprocess these assays remains an open research topic with substantial implications for harmonized atlases of cell states and fates. Moreover, the majority of single-cell transcriptomic discoveries are evaluated against flow cytometric and functional characterization.

Here we present a comparative evaluation of in silico and flow cytometric gating approaches for analyzing CITE-seq data. We investigate the relative strengths of decontPro and dsb as decontamination tools, and employ the scGate package to simulate in silico gating to allow interpretation of the downstream consequences.

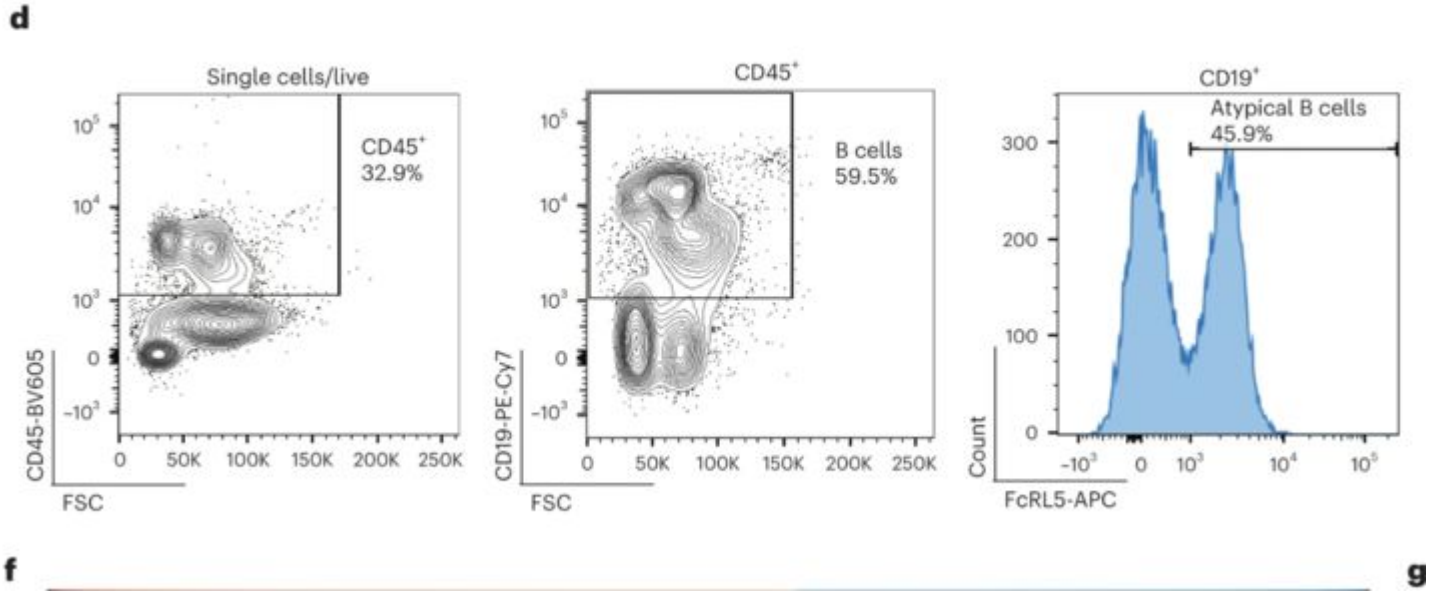
Importantly, when isotype controls and mRNA UMI counts are available, conclusions can be substantially affected by decisions to use or ignore these modalities in normalization, decontamination, and clustering.

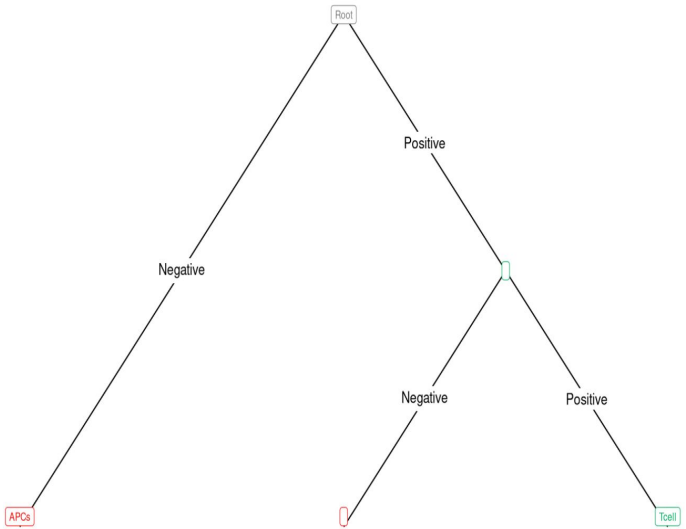
DecontX (which implements DecontPro) is a Bioconductor package that identifies and removes potential cell doublets and contaminating cells from single-cell data.

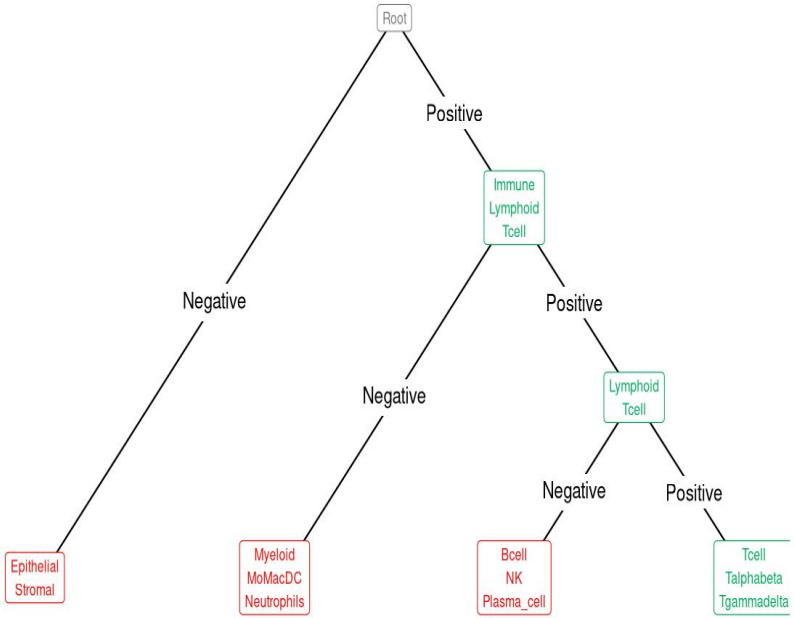
DSB, hosted on CRAN is another package that normalizes and denoises antibody derived tag data from CITE-seq datasets, and pioneered the use of isotype controls for background normalization.

scGate (hosted on CRAN) employs the UCell Bioconductor package to enable a reproducible, semi-supervised, in silico gating approach akin to more traditional flow cytometric gating. In conjunction with contemporary preprocessing and clustering-based workflows for CITE-seq data, scGate allows us to compare the outcomes of in silico gating on properly preprocessed CITE-seq data against flow cytometric counts of cells prepared via enrichment protocols. This provides a lens to judge the relative merits of decontamination workflows. Finally, we apply our findings from the above benchmarking experiments to a primary dataset of human bone marrow samples from healthy donors and pediatric leukemia patients. The results hold implications for clinical translation of multimodal single-cell profiling of patients and extensions to patient care in high-risk applications with no standard of care.

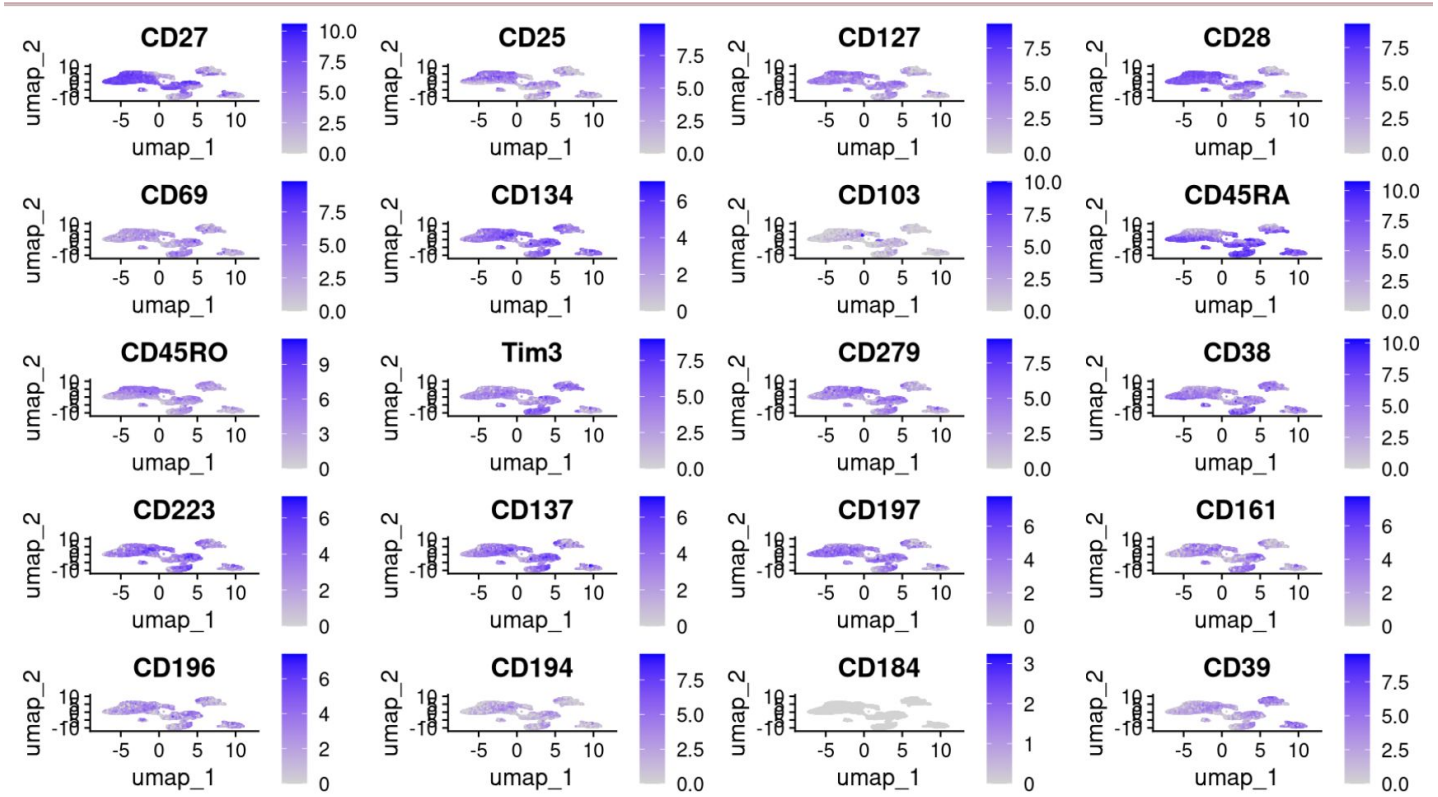
- Flow results from flow sorting B cells in https://www.nature.com/articles/s43018-022-00480-0/figures/3
-



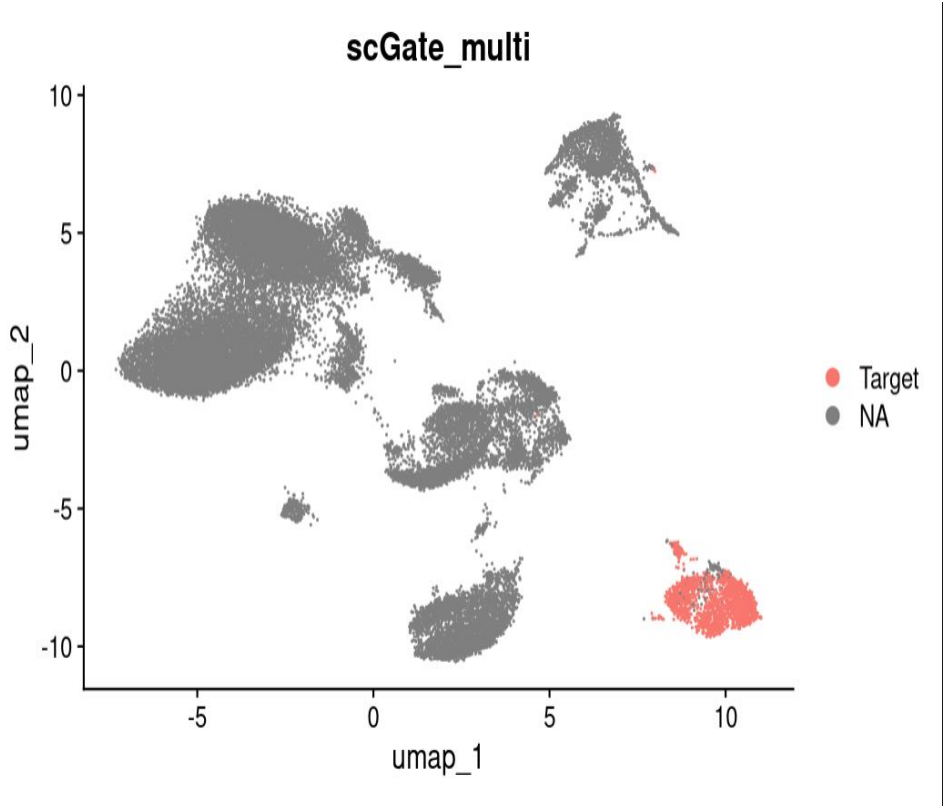

Root
Positive
Negative
Negative
Positive
APCs
Tcell

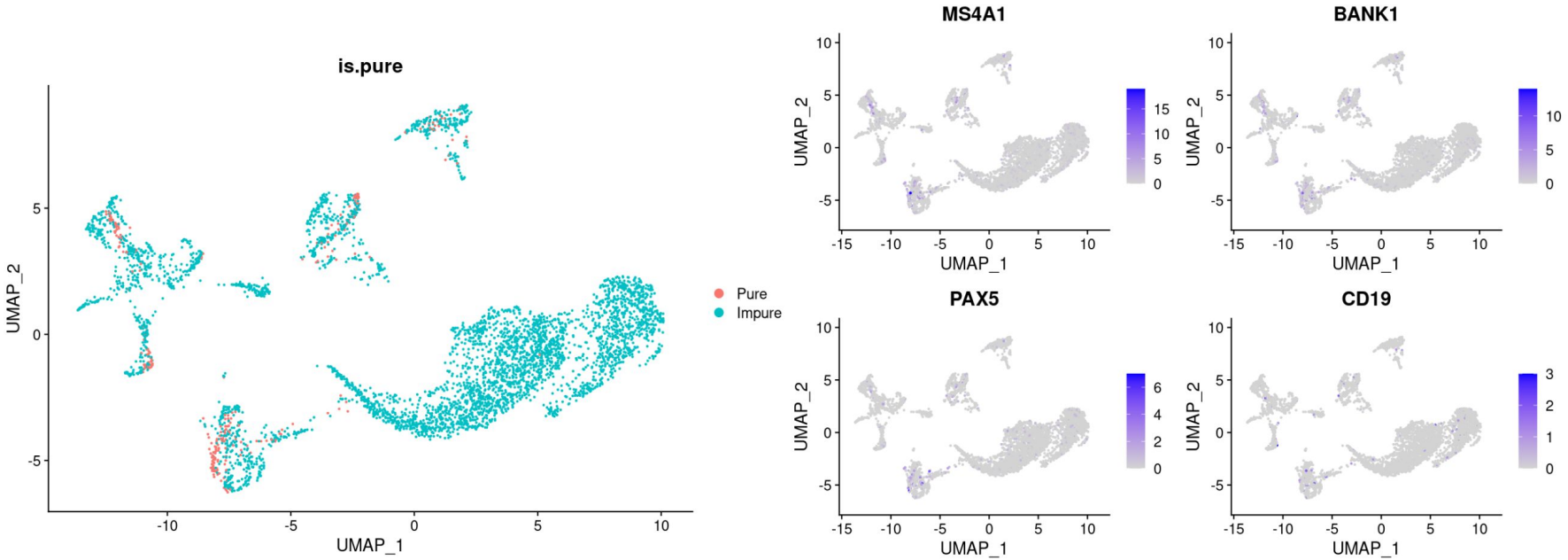

Root
Positive
Immune
Lymphoid
Tcell
Negative
Positive
Negative
Lymphoid
Tcell
Negative
Positive
Epithelial
Stromal
Myeloid
MoMacDC
Neutrophils
Bcell
NK
Plasma_cell
Tcell
Talphabeta
Tgammadelta

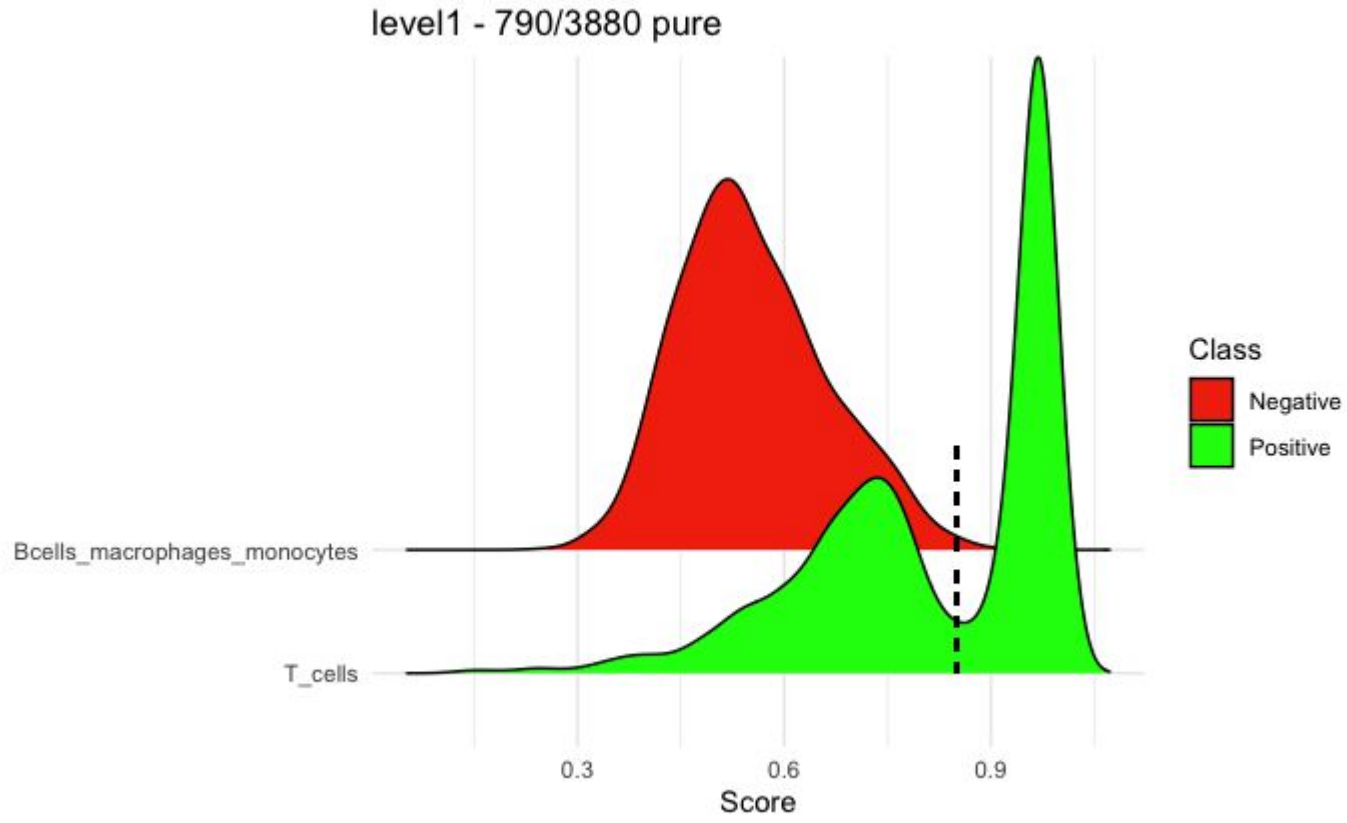
Expression of Positive Markers

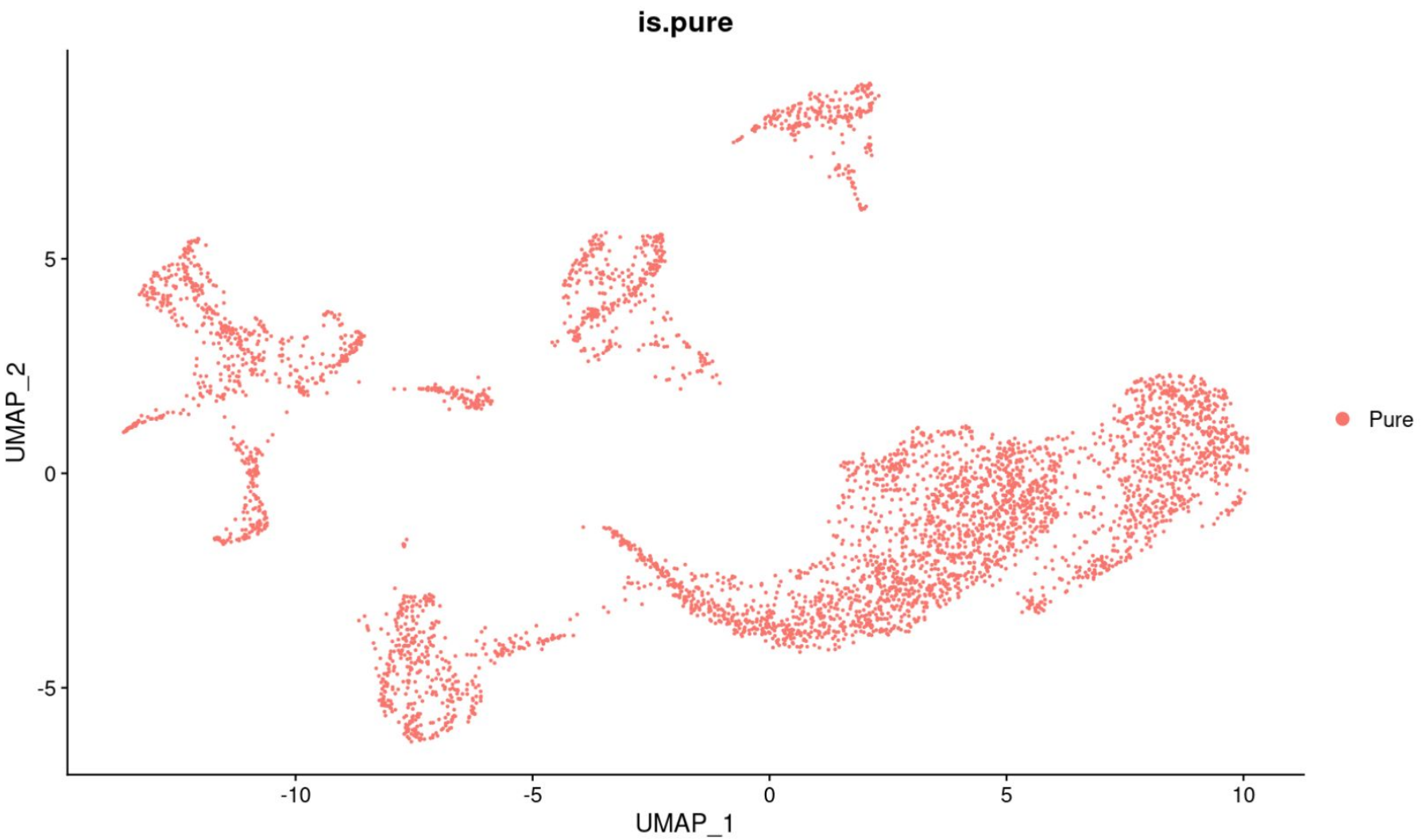


Labeled T Cells in Mair et.al After Gating




is.pure
UMAP_2
UMAP_1
Pure
Impure
MS4A1
BANK1
PAX5
CD19


level1 - 790/3880 pure
Bcells_macrophages_monocytes
T_cells
0.3
0.6
0.9
Score
Class
Negative
Positive


is.pure
UMAP_2
5
0
-5
-10
-5
0
5
10
UMAP_1
Pure