Supplementary Text 2: More Stringent Quality Control of OneK1K Cohort Single-cell RNA-seq Dataset

Laurie Rumker (Laurie_Rumker@hms.harvard.edu)
Joyce B. Kang (Joyce_Kang@hms.harvard.edu)
Soumya Raychaudhuri (soumya@broadinstitute.org)

April 11, 2023

1 Motivation

In collaboration with the authors of the OneK1K dataset index publication [1], we applied more stringent quality control to these PBMC scRNA-seq profiles before employing the dataset in our own analyses. Our additional dataset processing, summarized in this document, was prompted by our observations of isolated populations with mixed type assignments that expressed unexpected marker genes. We initially observed these putative doublet populations when performing a standard PCA-based analysis on each major cell type separately (e.g. PCA on cells labeled B cells). We observed fragmented cell populations with mixed type assignments (e.g. mixed B naive and B memory labels in a population separate from the major populations for these B cell subtypes) that also contained expression of unexpected marker genes that did not match the assigned labels (e.g. CD3 among these "B cells"). We we found that these populations corresponded to droplets identified as doublets by Demuxlet [2] or Scrublet [3] but not previously removed from the dataset.

2 Approach

We received scRNA-seq profiling (cells-by-counts matrix), as well as Demuxlet and Scrublet method output, directly from the study authors. Cell type assignments provided by the study authors were based on Azimuth mapping to a PBMC reference dataset [4]. After affirming basic per-profile QC thresholds were met (>200 genes, <8% mitochondrial gene reads), and removing 7680 genes that appeared in fewer than three profiles, we subdivided the profiles by major cell type using the following mapping from the 31 available labels to 7 major types:

- CD4+ T = [CD4 TCM, CD4 Naive, CD4 TEM, Treg, CD4 CTL, CD4 Proliferating]
- Other T = [CD8 TEM, CD8 Naive, CD8 TCM, MAIT, CD8 Proliferating, gdT, dnT]
- NK = [NK, NK_CD56bright, NK Proliferating, ILC]
- Monocyte = [CD14 Mono, CD16 Mono]
- DC = [cDC1, cDC2, pDC, ASDC]
- B = [B naive, B memory, B intermediate, Plasmablast]
- Other = [HSPC, Platelet, Eryth]

For each major cell type, we followed standard processing using scanpy (with parameters as described in the "Preprocessing and clustering 3k PBMCs" tutorial unless otherwise specified [5]) to total-count normalize to 10,000 reads per profile, logarithmize the data, retain only highly-variable genes and compute principal components (PCs). For each major cell type, we corrected these PCs for batch with harmony (batch = "pool", nclust = 50, sigma = 0.2, max_iter_harmony = 50) to generate hPCs. Resuming the scanpy pipeline, we used these hPCs to construct a nearest-neighbor graph and UMAP embedding per major cell type.

The index publication authors had previously removed any droplet identified as a doublet by both Scrublet and Demuxlet, but retained all droplets identified as doublets by only one of these two

methods. Of the 1,249,037 profiles provided by the OneK1K dataset authors, 22,662 were identified as doublets by Scrublet (predicted_doublet_mask==True) and 382,464 were called as doublets by Demuxlet ('BEST' assignment to 'DBL-'). We chose to remove these profiles. None of the cells included in the published dataset provided by the original authors had been classified by Demuxlet as ambiguous. Given that many profiles identified as doublets by Scrublet or Demuxlet were observed to cluster together transcriptionally in the dataset (Figures 1, 2, 3, 4, 5, 6, 7), we performed fine-grained clustering within each major cell type and removed any clusters for which >2/3 profiles were identified as doublets (by either Demuxlet or Scrublet).

We used Wilcoxon rank-sum tests to identify differentially-expressed genes per fine-grained cluster (scanpy's rank_gene_groups function with method = 'wilcoxon'). For major cell type groups besides "Other"—which contains the profiles assigned by Azimuth to the Platelet type—we also removed fine-grained clusters for which differential expression analysis identified PPBP, PF4, GP1BB and NRGN among the top 6 cluster-characteristic markers, suggestive of platelet doublets.

Finally, 1803 profiles lacked results from Demuxlet and Scrublet. Of these profiles, 131 were labeled "Doublet" by the publication authors and the remainder corresponded to an individual who also failed genotype data quality control in our analyses (not described here). We removed these 1083 profiles.

In summary, we removed each profile if:

- The profile was identified as a doublet by either Demuxlet or Scrublet OR
- The profile was assigned to a doublet-dominated fine-grained cluster OR
- The profile was labeled as a non-platelet type but assigned to a fine-grained cluster characterized by platelet-related genes OR
- The profile lacked doublet-calling results

Finally, we reassigned cell type labels to our retained cells, applying the same approach used by the publication authors for the initially-provided cell type labels: Azimuth reference mapping to the Azimuth PBMC reference. To accommodate Azimuth data volume limitations, we split the total dataset into 15 subsets by batch pool group and applied Azimuth separately to each subset. The major cell type classifications for the retained cells (i.e. among T, B, NK, and Myeloid groups) were unchanged for the vast majority of cells when compared to each cell's original major type assignment.

3 Results

Of the 1,249,037 profiles provided by the study authors from the published dataset, we chose to remove 416,556 (33%), the vast majority of which (405,126 profiles, 97%) were identified as doublets by either Scrublet or Demuxlet, and the remainder selected using our other two criteria (Tables 1, 2).

We found that the droplets identified as doublets by Scrublet or Demuxlet largely explained the isolated cell populations with mixed assigned types that we had observed, and platelet-contaminated populations explained some remaining fragmented populations (Figures 1, 2, 3, 4, 5, 6, 7, 8).

Major Type	Profiles	Resolution	Demuxlet	Scrublet	Fraction Removed
DC	6648	1.0	0.2	0.04	0.28
Mono	51876	2.0	0.22	0.02	0.25
B	129588	3.0	0.29	0.02	0.32
NK	172397	4.0	0.29	0.02	0.33
CD4+ T	624592	6.0	0.32	0.01	0.34
Other T	259893	6.0	0.31	0.03	0.34
Other	3912	0.2	0.44	0.04	0.61

Table 1: Profiles selected for removal, by major type. For each major type group, the total number of profiles assigned to that group ("Profiles") is shown, along with the resolution used for fine-grained clustering ("Resolution"), the fraction of all profiles identified as doublets by Demuxlet or Scrublet ("Demuxlet" and "Scrublet", respectively), and the fraction selected for removal based on all criteria ("Fraction Removed")

Removed	Scrublet Dblt.	Demuxlet Dblt.	Platelet Clust.	Dblt. Clust.	Count
F	F	F	F	F	832481
T	F	F	F	T	7734
T	F	F	T	F	1893
T	F	T	F	F	358161
T	F	T	F	T	23198
T	F	T	T	F	1105
T	T	F	F	F	18602
T	T	F	F	T	4033
T	T	F	T	F	27
T	NA	NA	NA	NA	1803

Table 2: Profiles selected for removal, by criterion. "Removed": T if the profile was selected for removal. "Scrublet Dblt.": T if Scrublet identified the profile as a doublet. "Demuxlet Dblt.": T if Demuxlet identified the profile as a doublet. "Platelet Clust.": T if the profile was assigned to a fine-grained cluster characterized by platelet-related genes. "Dblt. Clust.": T if the profile was assigned to a fine-grained cluster with <2/3 doublets. "Count": The number of profiles matching the combination of features captured by the corresponding row. The authors had previously removed profiles called as doublets by both Scrublet and Demuxlet. The final row captures profiles for which doublet-calling results were not available.

We make available a table containing the results of this data processing. This table contains one row per cell, indexed by barcode. In addition to cell-level metadata provided in the published dataset, we have added the following columns:

- demuxlet_DBL: True iff the cell was assigned as a doublet by Demuxlet
- demuxlet_AMB: True iff the cell was assigned as ambiguous by Demuxlet
- scrublet_DBL: True iff the cell was assigned as a doublet by Scrublet
- scrublet_score: Score assigned by Scrublet
- preQC_Azimuth_type: Azimuth-based cell types shared by the publication authors
- DBL_cluster: True iff the cell belonged to a cluster with >2/3 cells assigned as doublets by Scrublet or Demuxlet
- Platelet_cluster: True iff the cell belonged to a cluster characterized by platelet-associated marker genes
- remove_cellQC: True iff the cell met one of the four critera for removal described here
- remove_sampleQC: True iff the cell was associated with a sample we removed for our analyses (samples with low-quality or missing genotyping data, or labeled as ethnic outliers)
- fail_QC: True iff remove_cellQC or remove_sampleQC is True
- celltype: Azimuth cell type assignments for retained cells
- majortype: Major cell type assignments, aggregated from celltype

4 Discussion

Identification and removal of doublet droplets is a crucial quality control step in single-cell data analysis. Scrublet and Demuxlet are two of many available methods to accomplish this task. Scrublet simulates doublet transcriptional profiles as combinations of observed profiles and compares the observed profiles to these simulates. Demuxlet identifies droplets whose transcripts reflect a combination of genetic variants unlikely to arise from a single individual in the dataset. Because these methods have contrasting failure modes, applying both to the same dataset can enable the detection of droplets

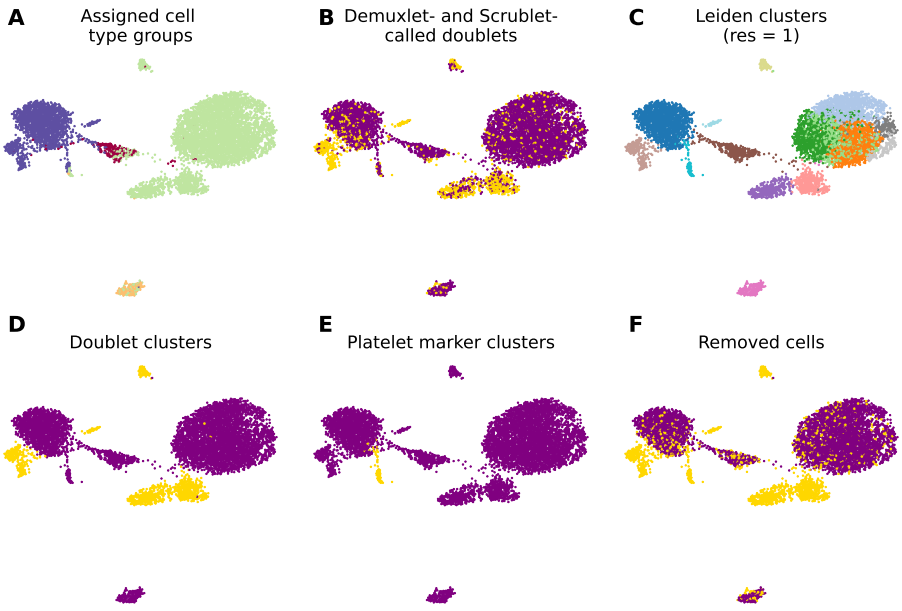

Figure 1: **Dendritic cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

by one method that were missed by the other. In the original publication of the OneK1K dataset, only cells called as doublets on the basis of both Scrublet and Demuxlet were removed, a small fraction of all identified doublets. Of the retained profiles identified as doublets, the vast majority were flagged by Demuxlet (i.e. on the basis of contrasting genotypes detected in the same droplet). We found that the retained doublets were transcriptionally perturbed in the dataset relative to cells identified as singlets and have chosen a more stringent quality control approach to remove these cells. In collaboration with the OneK1K dataset authors, we make available a table indicating which cells were selected for removal in our more stringent quality control.

References

[1] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, and Joseph E. Powell. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, April 2022.

[2] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, January 2018. Number: 1 Publisher: Nature Publishing Group.

[3] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4):281–291.e9, April 2019.

[4] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael

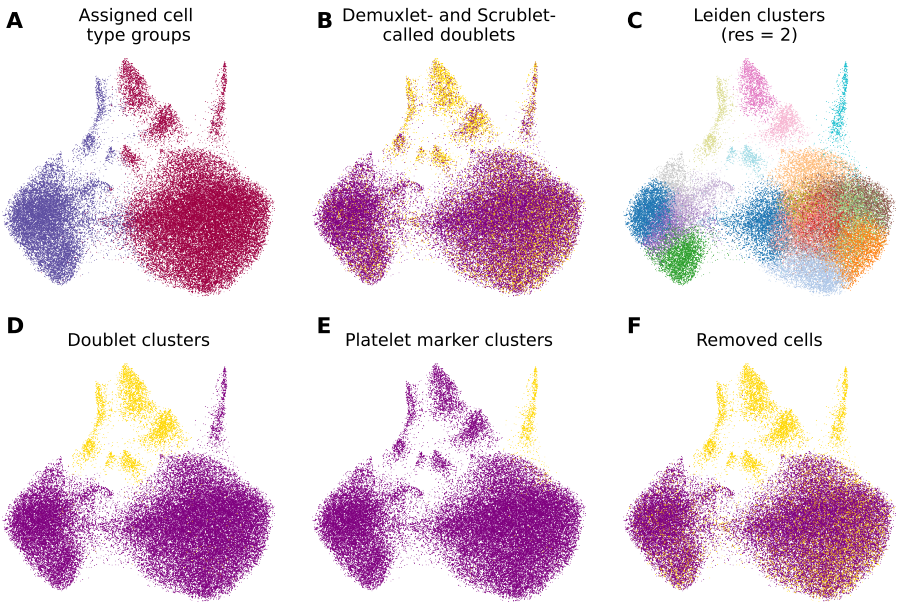

Figure 2: **Monocytes.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

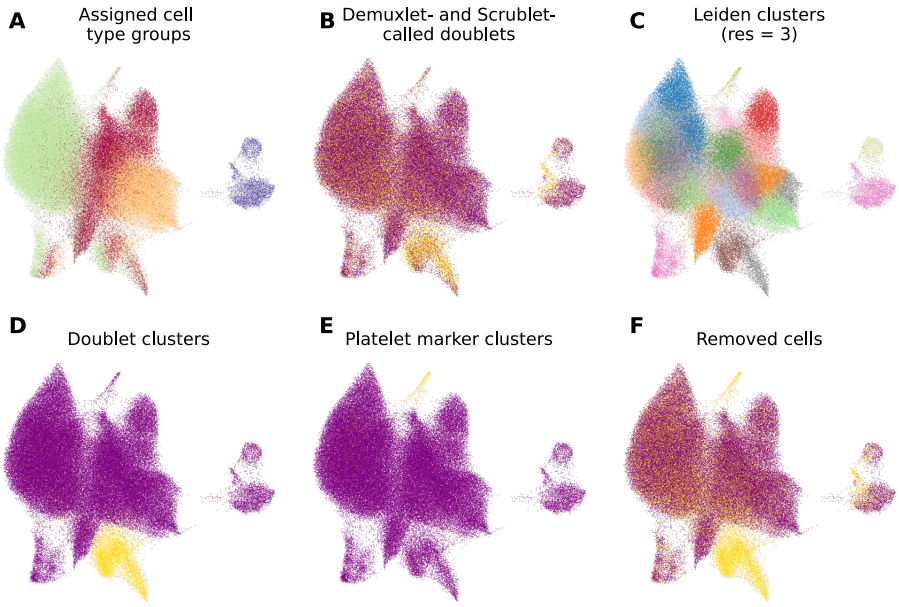

Figure 3: **B cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

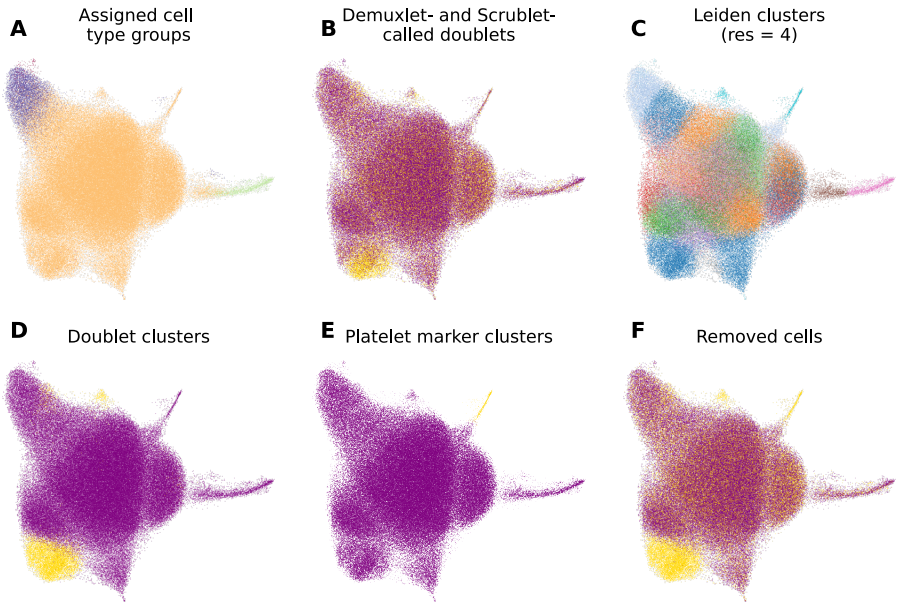

Figure 4: **NK cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

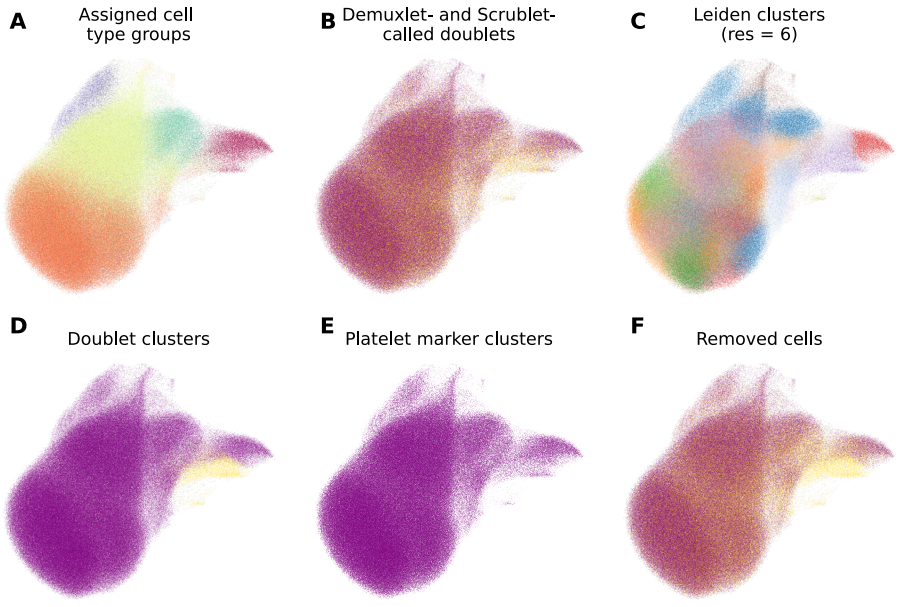

Figure 5: **CD4+ T cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

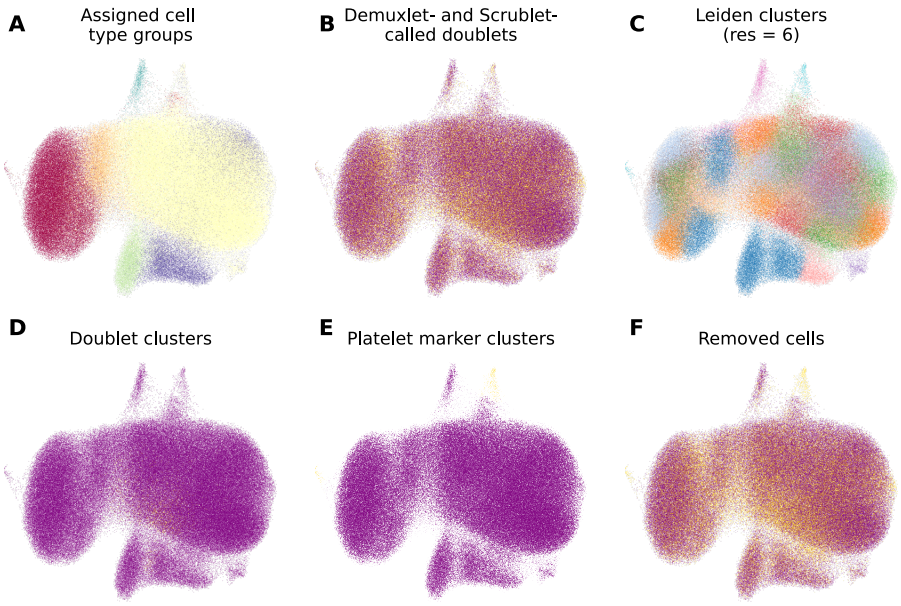

Figure 6: **Other T cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

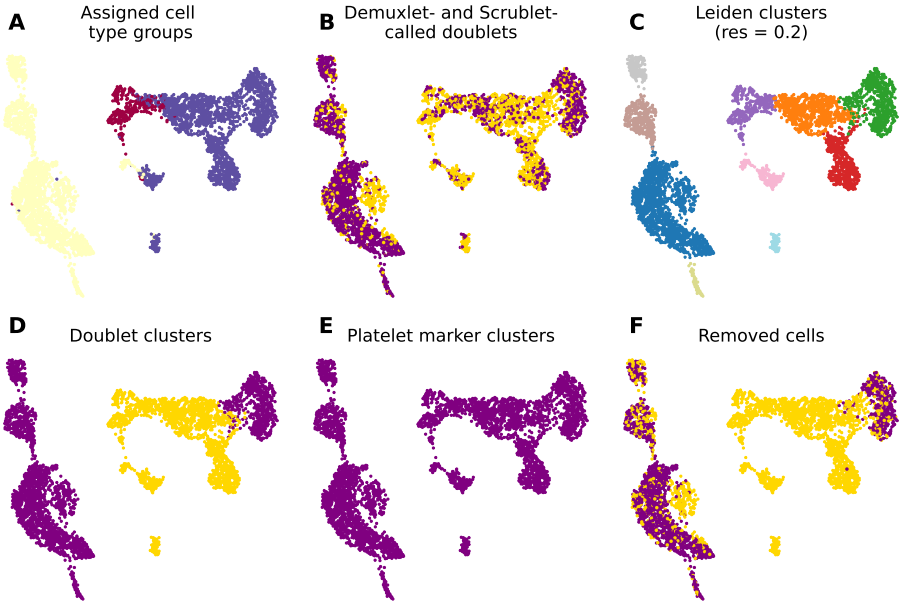

Figure 7: **All other cells.** (**A**) Profiles colored by type label, as provided by the publication authors. (**B**) Profiles identified by either Scrublet or Demuxlet as doublets, in gold. (**C**) Profile assignments to fine-grained clusters. (**D**) Clusters containing >2/3 profiles called as doublets by either Scrublet or Demuxlet, in gold. (**E**) Clusters characterized by platelet-related genes, in gold. (**F**) All profiles selected for removal, in gold; the union of gold profiles in B, D, and E.

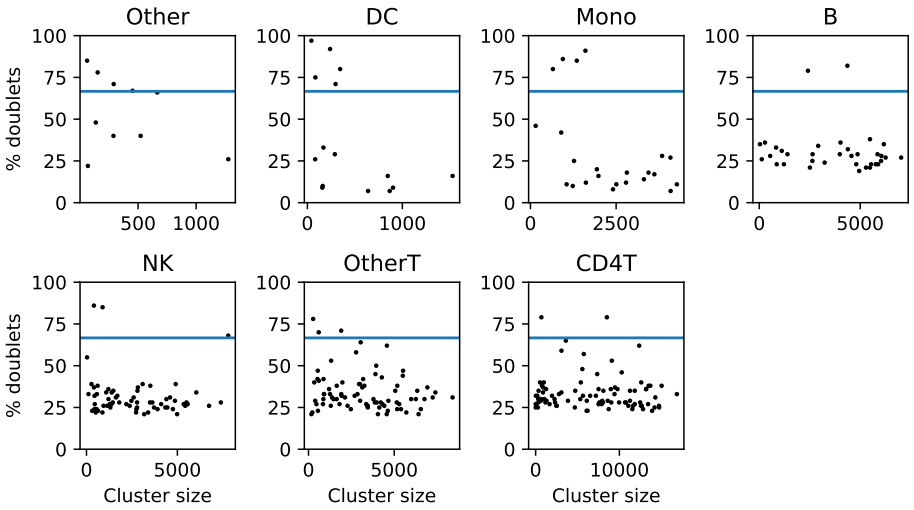

Figure 8: **Doublet cluster identification.** The profiles within each major type were clustered at a fine-grain resolution to identify and remove doublet-predominant clusters, in addition to isolated doublet profiles. The fraction of profiles called as doublets (by Scrublet or Demuxlet) for each fine-grained cluster is shown along the y axis, while the size of each cluster (number of profiles) is shown along the x axis, with plots separated by major type. Clusters with >2/3 doublets, above the blue line, were selected for removal.

Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. Publisher: Elsevier.

[5] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015.