Decontamination of ambient and margin noise in droplet-based single cell protein expression data with DecontPro

Yuan Yin[1], Masanao Yajima[2], and Joshua D. Campbell[1]

[1] Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA.
[2] Department of Mathematics and Statistics, Boston University, Boston, MA, USA.

Abstract

Assays such as CITE-seq can measure the abundance of cell surface proteins on individual cells using antibody derived tags (ADTs). However, many ADTs have high levels of background noise that can obfuscate down-stream analyses. We show a combination of two sources of contamination in the ADT data: 1) ambient ADTs from the cell suspension and 2) margin ADTs derived from the “spongelets” which are empty droplets with high levels of non-specific ADT expression. We propose a novel hierarchical Bayesian model, DecontPro, that can decontaminate ADT data by estimating and removing contamination from ambient and margin sources. DecontPro outperforms other decontamination tools in removing aberrantly expressed ADTs while retaining native ADTs and in improving clustering specificity after decontamination. Overall, DecontPro can be incorporated into CITE-seq workflows to improve the quality of downstream analyses.

Introduction

Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) is an assay that can quantify the abundance of RNA transcripts as well as cell surface proteins on individual cells[1]. Antibody-derived tags (ADTs) that bind to cell surface proteins are counted during sequencing to produce an ADT count matrix that quantifies the surface proteins. Information on the abundance of proteins complements single cell RNA-seq (scRNA-seq) data and makes the assay useful in cell type annotations, particularly for cell populations with more fuzzy transcriptomic profiles[2-4]. Other variants of CITE-seq with additional modalities and spatial information have been developed to understand complex biological processes[5, 6].

Previous studies have noted that individual ADTs often have a multimodal distribution including one lower "background" peak and one or more higher peaks attributed to the true signal from the cells. The lower background peak has been attributed to noise from non-specific binding of antibodies[7]. Algorithms such as TotalVI[8] have tried to leverage the multi-modal nature to identify and remove the lower background peak for each ADT. Other approaches try to measure and remove background levels using "spike-in" reference cells from another species such as mouse[1].

Contamination from various sources can contribute to poor-quality data in single-cell assays. In scRNA-seq data, ambient RNA from the cell suspension can be counted along with a cell's native RNA and result in contamination of gene markers between cell types[9, 10]. Ambient contamination may also occur in CITE-seq data as the methods for generating CITE-seq data also rely on microfluidic droplet-based devices. Two computational methods, dsb[11] and scAR[12], have been proposed that use the ADT expression profiles of the "empty droplets" to estimate and remove the noise from ambient material, where the empty droplets are droplets algorithmically labeled post hoc that are believed to contain no cells. However, these methods treat the empty droplets as a single source of noise. Furthermore, reliance on the empty droplets data may limit their application in cases where the empty droplets matrix is not available.

In this study, we analyzed four CITE-seq datasets and showed that there are three different types of "empty" droplets including 1) those mislabeled true cells with low RNA content, 2) those with low levels ADTs matching ambient distributions, and 3) those with medium levels of ADTs with non-specific distributions, which we denote as "spongelets". We also identified two contamination profiles, the ambient and the margin profile, from the ambient droplets and spongelets, that could each contribute to the ADT profiles in true cells and developed a novel Bayesian hierarchical model called DecontPro (**Decont**amination of **Pro**tein Expression Data) that estimates and removes both types of contamination. When applied to different ADT datasets, DecontPro was able to preserve the native markers in known cell types while removing contamination from the non-native markers. We demonstrate that DecontPro can outperform other tools in removing non-native markers and in improving downstream clustering. Finally, we show that DecontPro can increase the specificity of PD-1 expression in activated T and B-cells.

Results

Different contamination profiles contribute to CITE-seq data

We performed an exploratory analysis of CITE-seq and Total-seq datasets to understand heterogeneity among the droplets and to characterize different sources of contamination. We first analyzed a public dataset containing peripheral blood mononuclear cells from a healthy donor from 10x Genomics (PBMC 10K). Four distinct clusters of droplets were observed when comparing the total UMI counts of ADTs to total UMI counts of RNAs in each droplet (**Fig. 1A**). The "Cell" cluster (n=7,864 droplets) had high counts for both RNA and ADTs and were called true cells by Cell Ranger. The remaining three clusters were called empty droplets by Cell Ranger and had low RNA counts with varying levels of ADT counts. The Low Empty cluster (n=145,322 droplets) contained droplets with low levels of both total RNA and total ADT counts. The average profile of the droplets from this cluster was highly correlated with the average profile of droplets from the Cell cluster for both RNA and ADTs (R > 0.950 for all RNA and ADT datasets, except R = 0.769 for Golomb[13] ADT dataset; **Supplementary Figs. S7**) demonstrating that these droplets likely contain only ambient material[9].

The "High Empty" cluster (n=1,406) had an average of 5168 ADTs counts per droplet while the "Mid Empty" cluster (n=70,157) had an average of 268 ADTs. To understand the ADT profiles in the High-Empty and Mid-Empty clusters, we randomly sampled a thousand droplets from each of these two clusters and analyzed them with the droplets from the Cell cluster using the standard Seurat clustering workflow[14]. Droplets from the High Empty and Mid Empty clusters formed their own distinct clusters on the UMAP (**Fig. 1B**). The "High Empty" population had significantly higher levels of CD15 and CD16 compared to other populations, but few differentially expressed genes that identify a cell type (**Supplementary Figs. S1**). It likely represents neutrophils which are prevalent in PBMCs but have low RNA content[15, 16]. This suggests that viable cells with low RNA content may be readily characterized by ADT expression and that filtering of empty droplets should be performed separately for ADT and RNA data. In contrast to the High Empty cluster, the Mid Empty cluster did not show strong enrichment for any ADTs. The average profile of droplets in the Mid Empty cluster was not highly correlated to the average profile of the Low-Empty cluster (R = 0.312; **Fig. 1C**), confirming that the source of this cluster is not related to ambient ADTs.

The distribution of individual ADT expression in true cells is often multi-modal and contains more than one peak. For example, in the Cell cluster, the distribution of CD14, CD16, and CD45RA is bi-modal while the distribution of CD4 is tri-modal (**Fig. 1D**). The lower peaks have previously been characterized as background signals from non-specific binding of antibodies[7]. Interestingly, the density in the Mid Empty droplets largely overlapped with the lower peak of the Cell droplets density for the majority of ADTs (**Fig. 1D, Supplementary Fig. S3**). This was also observed in the three other ADT datasets (**Supplementary Figs. S4 – S6**). The average percentage of each ADT in the lower peak of the Cell cluster was highly correlated with the average percentage of each ADT in the Mid Empty cluster in the PBMC 10K dataset (R = 0.976; **Fig. 1E**), and to a lesser degree in two other datasets (R = 0.720 and R = 0.641; **Supplementary Figs. S9**). These findings suggest association between the lower "background" peaks in the Cell droplets and the Mid Empty droplets.

The low-level background expression is also manifested after normalizing Cell droplets by their library sizes and classifying them by their cell types. (**Supplementary Fig. S10**). Low peaks of ADTs were observed in cell types that are not supposed to express the ADTs. We call this low-level contamination of ADTs the margin profile. Interestingly, the density showing the margin

profile for each ADT has a consistent mean across cell types (**Supplementary Figs. S10 – S12**). Overlaying the ambient profile as a percentage attributed to each ADT, the two profiles show different means for most ADTs.

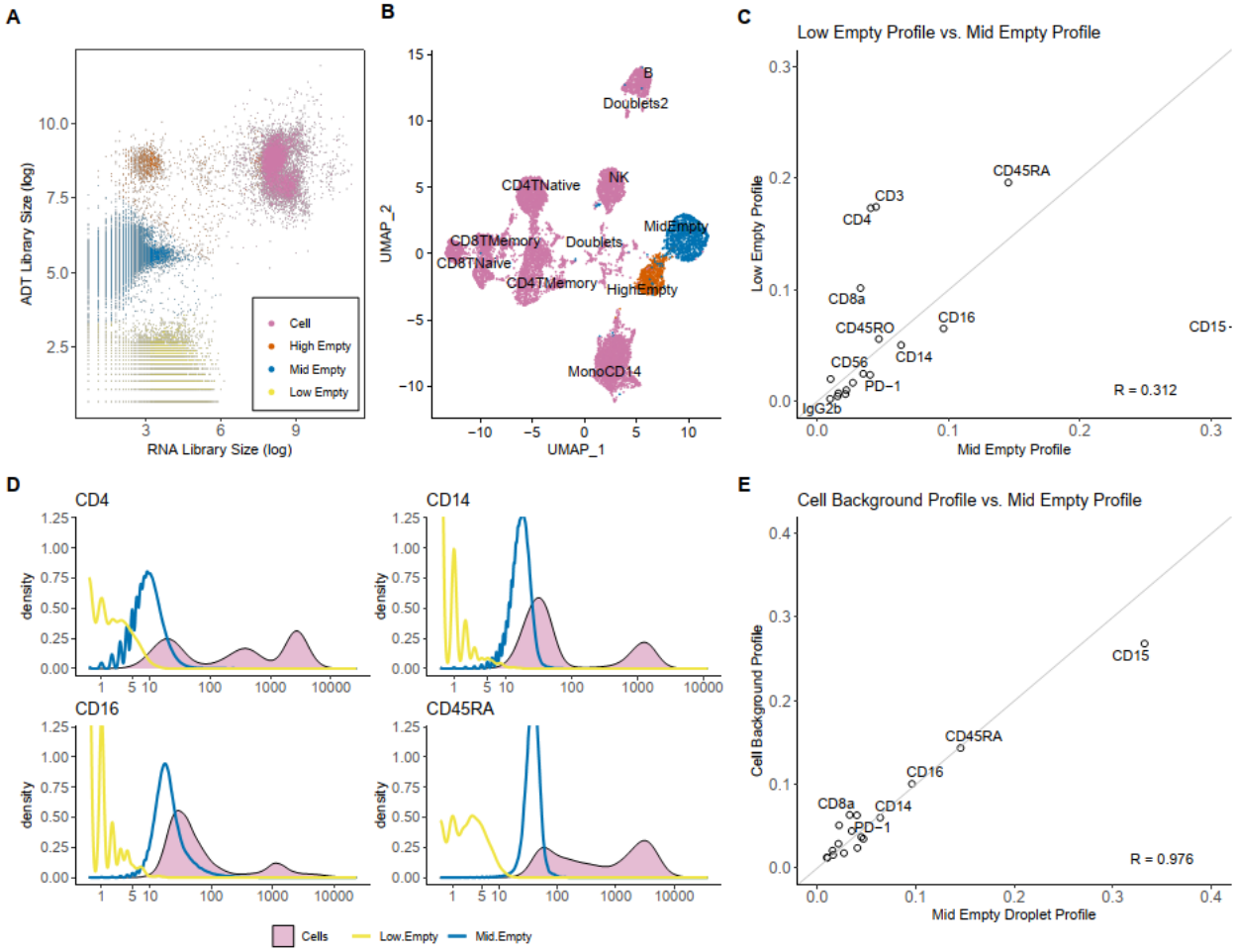


Figure 1. Analyses of ADT droplets in the PBMC 10K dataset. A. Among all droplets, only the cluster with high RNA-seq library size were called as cell droplets by the 10x Genomics software, while we observed clusters with varying levels of ADT library sizes among the droplets labeled as empty. We identified three empty droplet clusters, and called them High, Mid, and Low Empty clusters. **B.** A thousand droplets sampled from High and Mid Empty clusters each were combined with Cell droplets for clustering with Seurat. Distinct clusters were observed. **C.** The empirical distribution built using the Mid and Low Empty droplets showed poor correlation. **D.** Density plots of droplets from the Mid Empty cluster were close to the background peak of the Cell density plot. **E.** We manually curate the background peaks of the Cell density and normalize them into a multinomial distribution. The distribution was found to be highly correlated with the Mid Empty cluster empirical distribution.

A novel model for estimating and removing decontamination

We next sought to build a deconvolution algorithm that can estimate and remove contamination from the ambient profile and the margin profile. We assume that each cell is a mixture of three sources: 1) ADTs from the native cell, 2) ADTs from the ambient material present in the cell suspension, and 3) ADTs contributed by the margin profile (**Fig. 2A**).

The native ADT expression of a cell population k is characterized by a multinomial distribution φ_k, where φ_{kj} represents the percentage of native expression attributed to ADT j. The ambient profile is characterized by a multinomial distribution η where η_j represents the percentage of ambient expression attributed to ADT j. We use a mixing parameter θ_{ij} to model the proportion of counts contributed by the ambient ADT j to the droplet i, and β_{ij} to model the proportion of margin ADT j in droplet i. We use parameters μ_j as the prior means for β_{ij} and parameter δ_i as the prior means for θ_{ij}. Specifically, μ_j represents the mean margin level for ADT j, and δ_i represents the average ambient contamination level in droplet i. After mixing the parameters and scaling by the library size, the observed counts are generated through a Poisson distribution. We use the variational inference framework provided by Stan to make inference on the parameters, and deconvolute the ADT count matrix into the Native, Ambient and Margin matrices for downstream analysis (**Fig. 2B**). Detailed description of the model can be found in the **Methods** section.

To model the ambient profile η in a dataset, we calculate the average of the ADT across droplets that pass filtering. In many scRNA-seq datasets, the average cell profile is highly correlated to the empirical distribution of the ambient droplets. We also observe this pattern in three ADT datasets (PBMC 10k: R = 0.977, PBMC 5k: R = 0.991, MALT 10k: R = 0.968; **Supplementary Fig. S7**). The correlation is less strong for one dataset (Golomb: R = 0.769), primarily due to Ly6C which was among the more highly stained ADTs in the dataset. Visualizing the staining level on a UMAP, Ly6C also had a high staining level globally across droplets, in contrast to other highly stained ADTs with staining more localized (**Supplementary Fig. S8**). This suggests additional technical noise in this dataset that might be related to the experiment setup, confounding the correlation. Alternatively, when empty droplets matrix is available and is provided, our software package can estimate the ambient profile η using the empirical distribution of the ambient droplets, which should be a closer approximation to the true η.

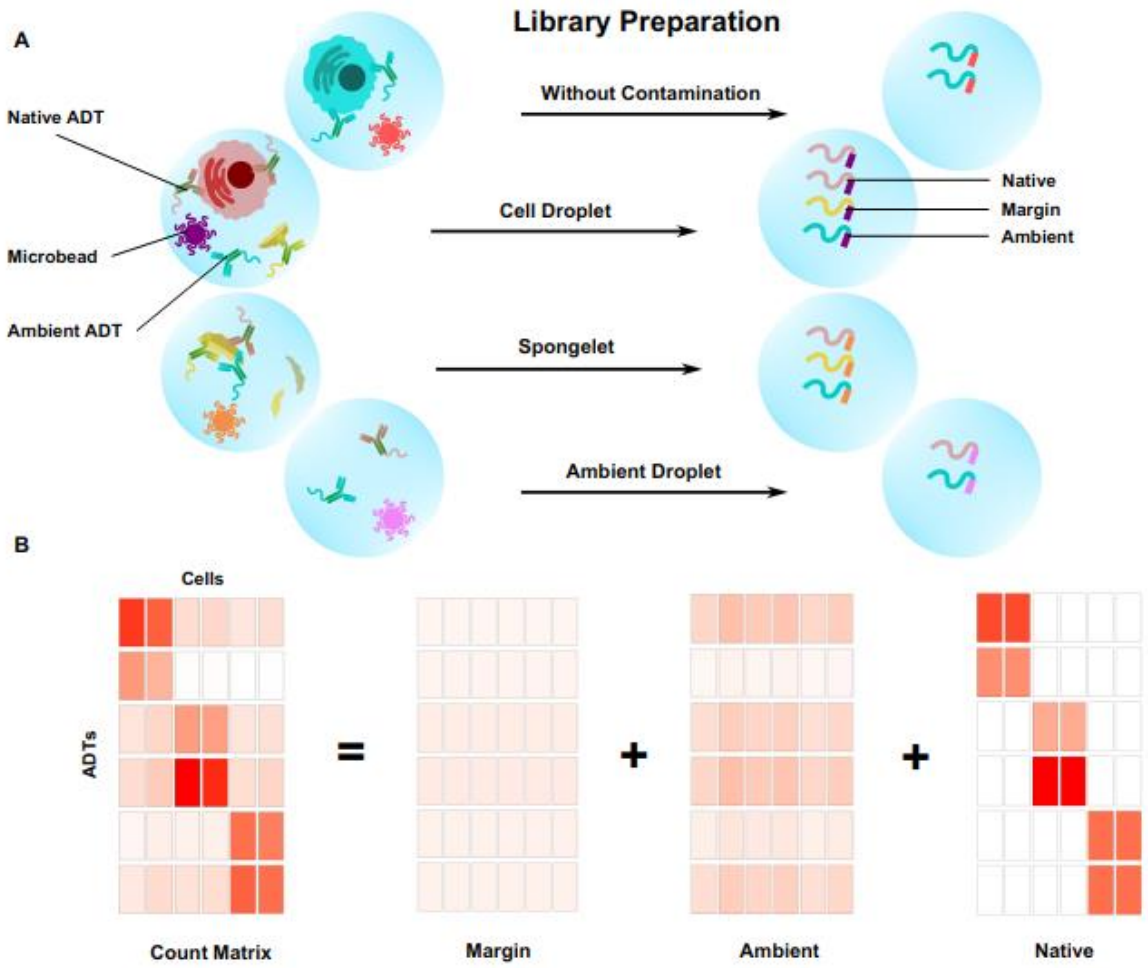


Figure 2. Visualizing the model. A. Single cells and ADTs are isolated in droplet compartments. Native ADTs refer to those that bind to their target cell surface proteins, whereas ambient ADTs are those present in the cell suspension. Spongelets contain some non-specific distribution and an intermediary level of ADTs, while ambient droplets have a low level of ADTs attributed to the ambient source. These sources of ADTs can contribute to the cell droplets as they are pooled together during library preparation to form the droplet libraries to be sequenced. **B.** In our model, we describe the profile observed in the spongelets as the margin profile and that in ambient droplets as the ambient profile. After decontamination, the count matrix of features-by-cells is decomposed into three matrices: margin matrix, ambient matrix, and native matrix, with each containing a portion of the counts that are attributed to the respective sources.

Decontamination Results of PBMC 10k

To demonstrate the ability of our method to improve data quality, we applied DecontPro to the PBMC 10K dataset from 10X Genomics and compared the distributions of counts before and after decontamination. In the original count matrix, all ADTs were expressed to some degree in nearly every cell cluster (**Fig. 3A**). After decontamination, expression markers were left for cell types (**Fig. 3B**). For example, CD19 is a B-cell marker with an average expression of 668.46 in the B-cell cluster and a low average expression of 13.68 across other cell types in the original count matrix. After decontamination the only clusters retaining expression of CD19 were the B-cells and doublets. Similarly, T-cell marker CD3 had remaining counts only in the T-cell clusters and a doublet cluster.

We superimposed the density of the estimated margin level onto the density of the ADTs after normalizing by the droplet library sizes and found that the estimated margin largely overlapped with the low-level peak across cell clusters (**Fig. 3C**). As expected, native ADTs in their corresponding clusters were expressed at a higher level, and hence had a higher peak in the normalized counts' density. For example, CD3 had a peak near 0.15 in the CD8 Naive T-cells cluster, indicating that about 15% of all counts in the cluster were attributed to CD3. In this case, the CD3 margin was estimated using other cell clusters, particularly those that did not have CD3 as their native ADT, such as NK cell clusters. The estimated margin had a small variance in general, in contrast to some normalized count densities. CD16 in the CD16 Monocyte cluster had a near-zero background peak, but also had a tail that extended beyond 0.15. This may be due to cells in the cluster that expressed CD16 surface protein natively. Our conservative estimation of margin ensured that we did not mistake native counts as margin counts, particularly for droplets that had a low native count.

To systematically assess how well the algorithm performed in specificity in retaining native ADTs for clusters, we calculated the percentage of cells in each cluster expressing marker ADTs before and after decontamination (**Fig 3D**). Markers and cell types included CD3 and CD4 for CD4+ T-cells, CD3 and CD8 for CD8+ T-cells, CD19 for B-cells, CD14 for CD14+ monocytes, CD16 for CD16+ monocytes, and CD56 for NK cells. In many cases, the algorithm was able to greatly reduce or completely remove aberrant counts in non-native cell types. For example, CD14 expression was removed from T-cells, CD19 was removed from T-cells, NK-cells, and monocytes, and CD3 was removed from B-cells and monocytes. In some cases, the reduction left some ADT values in unexpected cell clusters. The monocyte marker CD14 was still detected in a high percentage of B-cells after decontamination (98%). It is noteworthy that the threshold for detecting a marker is one count. Therefore, even though the marker is detected in a high percentage of B-cells, the overall levels of CD14 in B-cells was still greatly reduced compared to the original ADT counts (**Supplementary Fig. S13**).

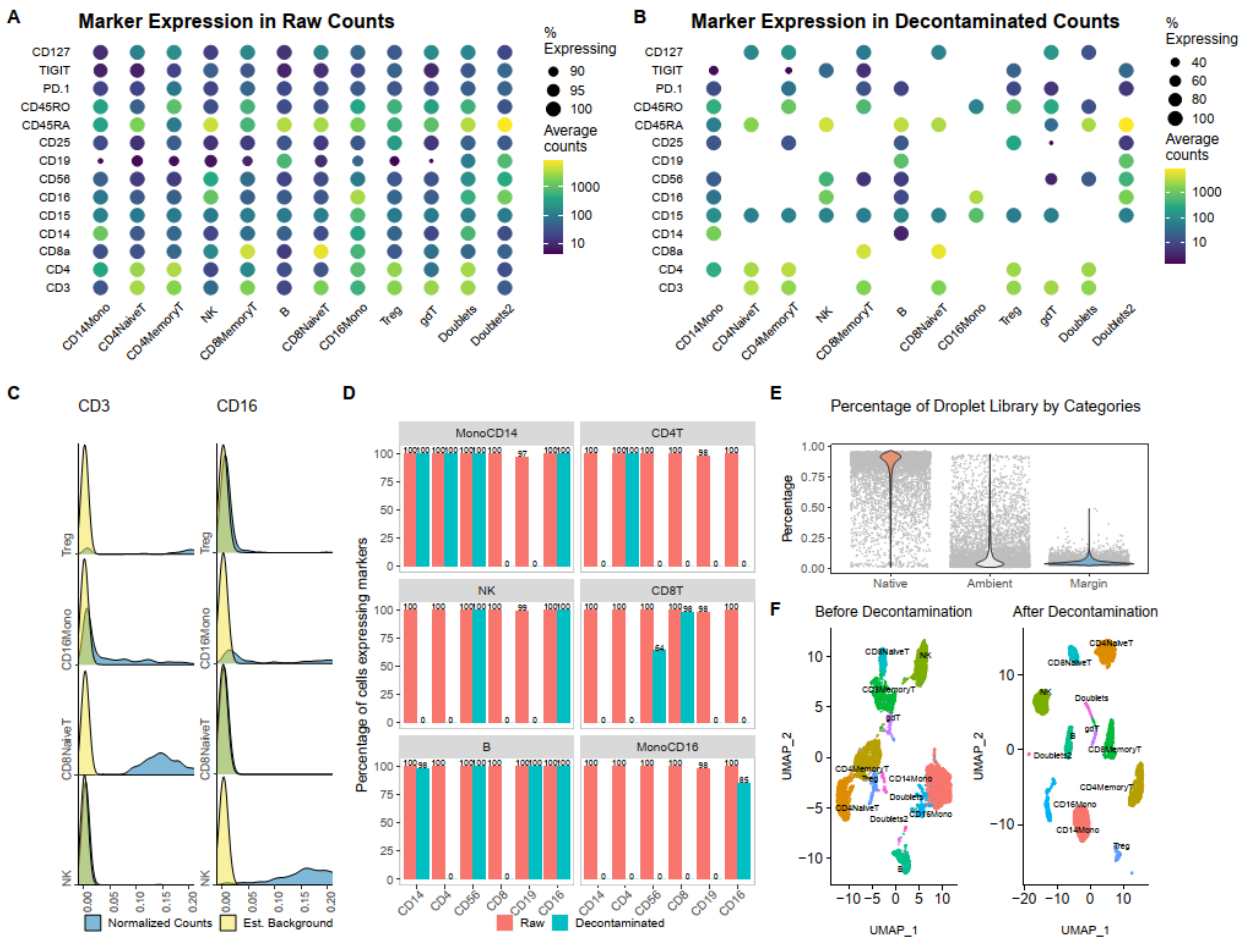


Figure 3. Decontamination of the PBMC 10K dataset with DecontPro. Percentage of droplets in each cell cluster that were expressing the marker ADTs and the cluster-wise average expression level **A.** before decontamination and **B.** after decontamination. **C.** Percentage of cells in clusters being identified as T-cells (CD4, CD8), B-cells (CD19) Monocytes (CD14, CD16), and NK cells (CD56) that were expressing these cell type markers before and after decontamination. Markers were deemed to be expressed when their count in a droplet was more than one. **D.** Density plot of selected ADTs after normalizing by droplet library sizes. The model estimated background was superimposed onto the density plot. **E.** Model estimated native, ambient and background counts of droplets were converted into a percentage of total library counts of droplets. **F.** Decontaminated counts showed improved clustering results. Cluster labels in new clustering results were from the original clusters.

DecontPro also estimates the percentage of counts contributed by the native, ambient and margin signals (**Fig. 3E**). As expected, the native counts took the highest percentage of libraries across droplets on average (median 89.5%), whereas the margin counts were estimated to be consistently at a basal level (median 4.4%). Compared to the background, ambient counts percentage (median 5.5%) estimation have noticeably more outliers. UMAPs generated with the decontaminated counts displayed more separate clusters compared to the UMAP generated with the original counts (**Fig. 3F**). Droplets that were between cell clusters on the UMAP before were closer to cluster centers after decontamination, with clusters showing clearer separations, indicating an improvement in clustering quality.

Benchmarking against Other Methods

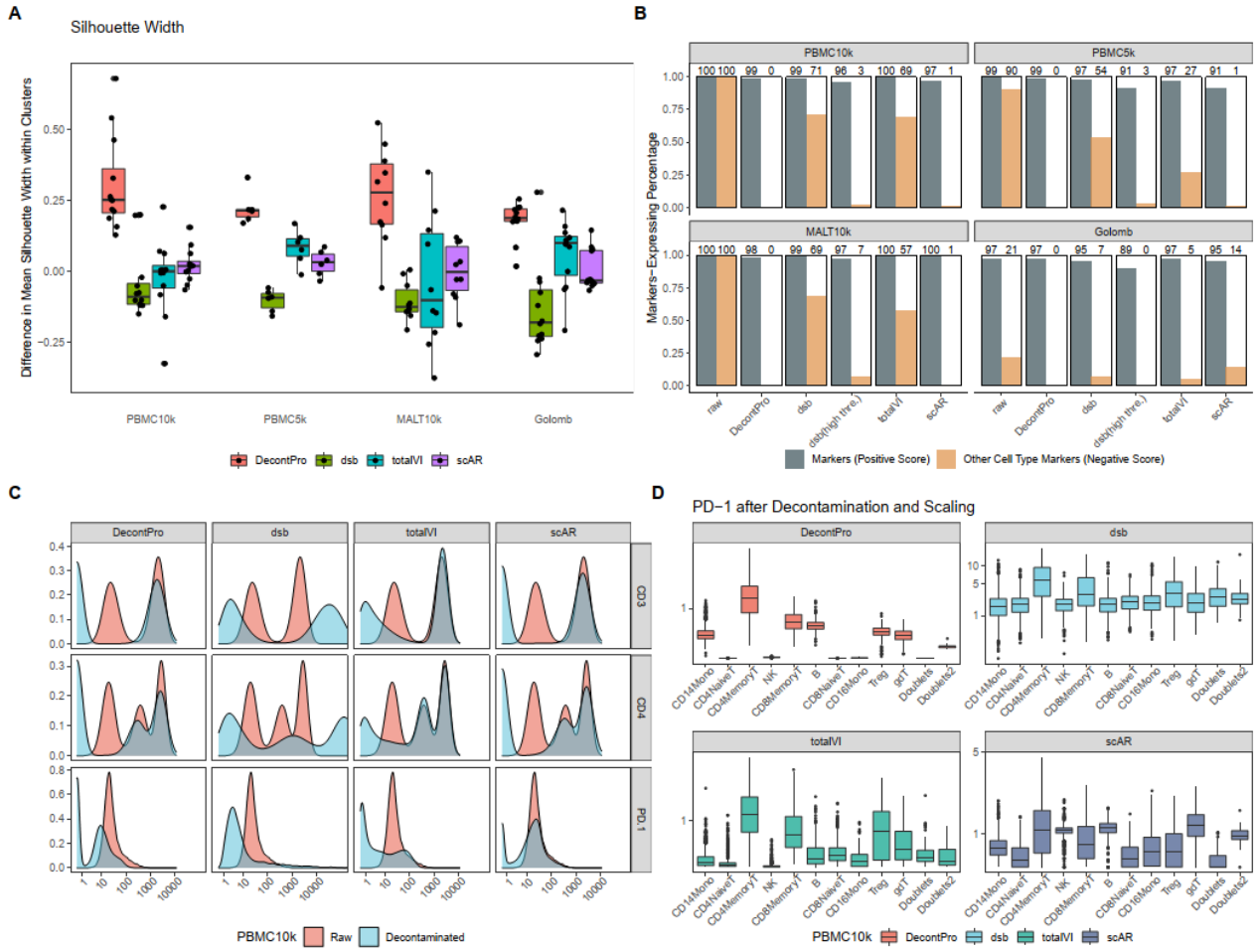


Figure 4. Benchmarking of ADT decontamination methods. We compared the ability of four algorithms to remove contamination from four datasets. The other methods include dsb, scAR, and totalVI. **A.** The average silhouette width for each cell cluster identified by Seurat was calculated before and after decontamination of ADT counts. A higher difference in silhouette widths indicates better improvement in cell similarity within that cluster. On average, DecontPro improved silhouette widths better than other methods for each dataset. **B.** For each dataset, a pair of scores were calculated after applying each decontamination algorithm to ascertain the degree to which each algorithm maintains true expression while removing contamination. The "positive" score was defined as the percentage of cells expressing native markers averaged across cell clusters. In contrast, the "negative" score was defined as the percentage of cells expressing non-native markers averaged across cell clusters. In the CD4 naïve T-cells cluster for example, CD4 and CD45RA were defined as the native markers, while CD8 and CD45RO were defined as the non-native markers. As dsb outputs normalized continuous counts after decontamination, we applied two different thresholds to determine which markers were detected in each cell. Overall, DecontPro performs the best in having high native scores and low non-native scores. **C.** Density plot of CD3, CD4 and PD-1 before and after decontamination in the PBMC 10k dataset. The dsb output was exponentiated to allow for comparison with the raw count density. **D.** Box plot of PD-1 level in cell clusters after decontamination and scaling PBMC 10k dataset.

We benchmarked DecontPro against three other decontamination algorithms applicable to CITE-seq datasets: dsb, scAR, and totalVI. To compare how well decontamination improved clustering, we calculated the mean silhouette width of each cell cluster before and after decontaminating four public datasets. Higher silhouette widths indicate higher similarities between cells within each

cluster. Across all datasets, our algorithm showed the highest average improvement in silhouette widths after decontamination (**Fig. 4A**). We next sought to understand how effectively the algorithms improved the specificity of marker expressions in cell clusters. Specifically, we wanted to measure if the expected markers in annotated clusters retained their expression, while the unexpected markers were removed. We calculated two scores for each algorithm on each dataset. A positive score which calculated the percentage of droplets in the clusters that expressed the expected markers, and a negative score calculated the percentage of droplets in the clusters that expressed unexpected markers. An effective algorithm would score close to 100 for positive scores and close to 0 for negative scores. The full list of annotated clusters and markers used can be found in the **Methods** section. Raw data had both scores high, and for PBMC 10k, PBMC 5k and MALT 10k dataset the negative scores were close to a hundred, indicating a high level of contamination (**Fig. 4B**). When using the difference between positive score and negative score to measure the algorithm performance, DecontPro showed high overall score, indicating that it retained the markers while removing the unexpected markers across datasets (PBMC 10k: 99, PBMC 5k: 99, MALT 10k: 98, Golomb: 97). As dsb performs its own normalization and scaling, the decontaminated counts are not directly comparable with the raw count matrix. We therefore used two different thresholds for a maker to be considered detected in a droplet (dsb: 1 count; dsb high threshold: 5 counts) and calculated the scores accordingly. The higher threshold for dsb produced better negative scores more similar to algorithms such as DecontPro and scAR, while slightly compromising the positive score in each dataset. Overall, the scores suggested a better decontamination results when using high threshold (PBMC 10k: 93, PBMC 5k: 88, MALT 10k: 90, Golomb: 89) than using low threshold (PBMC 10k: 28, PBMC 5k: 53, MALT 10k: 30, Golomb: 88). It is also notable that scAR performed well in this measure (PBMC 10k: 96, PBMC 5k: 90, MALT 10k: 99, Golomb: 81).

For the PBMC 10k dataset, the decontamination results for known cell type markers were mostly similar across algorithms. For antibodies with high expression, such as CD3 and CD4, all algorithms retained the higher peak while significantly reducing the lower background peak (**Fig. 4C**). However, PD-1 was expressed at lower levels and demonstrated a noticeable difference in decontamination results between algorithms. PD-1 is known to be predominantly expressed in activated T-cells, along with B-cells, natural killer (NK) cells and myeloid cells[17-19]. Our algorithm reduced the PD-1 expression significantly in the naïve T-cells population, while preserving the expression level in the CD4 memory T-cells, CD8 memory T-cells, and Treg clusters where we expected a high PD-1 level (**Fig. 4D**). Besides, some levels in the B-cells and monocytes clusters were also retained. Dsb and totalVI largely repressed PD-1 expression in all clusters except CD4 memory T-cells, CD8 memory T-cells, and Treg cluster. ScAR is the only algorithm that retained PD-1 in the NK cells cluster but has also removed its expression in the Treg cluster.

Discussion

We developed DecontPro, a Bayesian statistical model, that decontaminates two sources of contaminations we observed empirically in CITE-seq data. Our analyses suggested that the empty droplets, which are often overlooked in single cell data analysis or are thought to homogeneously possess ambient noise, are heterogeneous with signals connected to different sources. In particular, we identified clusters of empty droplets with higher levels of ADTs with different distributions from that of the ambient droplets, and we referred to them as “spongelets”. This observation motivated us to design a decontaminating algorithm that decomposes the ADT count matrix into three components: native, ambient and margin, with the native representing the contribution from the cells, and the ambient and margin representing the contribution from the ambient droplets and spongelets.

DecontPro works in situations when empty droplets are not available. Even though we showed that empty droplets possess valuable information about sources of contaminations, a motivation of our model design is to allow decontamination using only cell-containing droplets. The output from the 10x Genomics pipeline contains both empty and cell-containing droplets, and the number of empty droplets can be hundreds of folds of the number of cell-containing droplets. Despite the sparsity of the empty droplet datasets, they can still pose a computational challenge for benchtop computers with limited memory. In addition, these empty droplet datasets are usually not uploaded to the public repositories when researchers share their datasets. Therefore, we seek to approximate the contamination signals in empty droplets using the cell-containing droplets. We observed a strong correlation between the cell-average distribution and the empirical distribution of low empty droplets, which represent the ambient noise, in three out of the four datasets we analyzed. We also investigated the one dataset where the correlation is less strong and found it to have an ADT with an aberrant staining level. The hypothesis underlying the similarity between the cell-average distribution and ambient distribution is that the ambient molecules are contributed by each cell equally likely, resulting in the ambient distribution being the average distribution of cell distributions. However, it comes with caveats when there are technical procedures such as cell sorting or over-staining of ADTs that could alter the cell type proportion or cell distributions. In this case, our software allows users to input low empty droplets for a more accurate estimation of the ambient signal.

Our hierarchical Bayesian model is flexible and tunable. The hierarchical structure allows for variation in individual droplets and individual ADTs to be modeled when estimating contaminations. The prior parameters that influence the variation can be tuned by the user to account for more “spread-out” contamination levels. This would alter the percentage of droplet libraries that can be explained by contamination components and in effect tunes the aggressiveness of the model in decontaminating. Diagnostic plots in the software package help the users to evaluate if decontamination is done appropriately. For example, in our analysis of the PBMC 10k dataset, prior knowledge of PD-1 was used to check the decontamination results and affirmed that the prior parameters were set reasonably. We also allow the ambient distribution to be supplied to the model in case users have access to the empty droplet dataset or are confident that a certain distribution represents the ambient distribution better.

Methods

Empty droplet analyses

PBMC 10k dataset with a total of 6,794,880 raw droplets and 7865 filtered droplets, was downloaded from the 10x Genomics website. Droplets with zero total counts for both ADT and RNA were filtered. The library sizes for ADT and RNA were calculated by summing all the ADT and RNA counts in each droplet. The droplets with barcodes in the raw but not filtered dataset were named empty droplets. High, Mid and Low Empty clusters in empty droplets were identified using k-means with the number of clusters set to 3. Mid and Low Empty profiles were calculated by summing ADT counts over droplets and normalizing the vector to one. The cell background profile was calculated by summing ADT counts over droplets that fall into the background peak and normalizing. The background peaks are manually identified using the bounds shown in **Supplementary Figures S14-S16**. Cell clusters were identified with markers: MonoCD14: (CD14+), CD4TMemory: (CD3+, CD4+, CD45RO+), CD4TNative: (CD3+, CD4+, CD45RA+), CD8TMemory: (CD3+, CD8+, CD45RO+), NK: (CD16+, CD56+), B: (CD19+), CD8TNaive: (CD3+, CD8+, CD45RA+).

Data Generation Model

For a droplet i containing a cell of type k[i] having the library size L_i, we assume the jth ADT count x_{ij} is a realization from a Poisson distribution with rate parameter λ_{ij}.

$$x_{ij}|L \sim Poisson(\lambda_{ij} L_i)$$

When there is a total of I droplets, J ADTs and K cell types the count matrix can be organized into an I by J matrix **X**=[x_{ij}] and a corresponding cell type indicator vector **k**=[k[i]].

We will make further assumptions that the rate parameter λ_{ij} can be deconvolved into three components with:

1. A simplex for native profile of cell type k[i]: $\phi_{k[i]}$,
2. A simplex for ambient contamination profile: η,
3. A margin proportion of counts due to non-specific binding: β_{ij}

After subtracting the counts due to non-specific binding, contamination due to the ambient profile will be expressed using θ_{ij} hence

$$\lambda_{ij} = \left(1 - \beta_{ij}\right) * \theta_{ij} * \eta_j + \left(1 - \beta_{ij}\right) * (1 - \theta_{ij}) * \phi_{k[i]j} + \beta_{ij}$$

To specify the prior distributions, for each ADT j, we use μ_j to model the mean margin proportion across droplets. It serves as the prior mean for β_{ij} which follows a truncated normal distribution bounded by (0, 0.5).

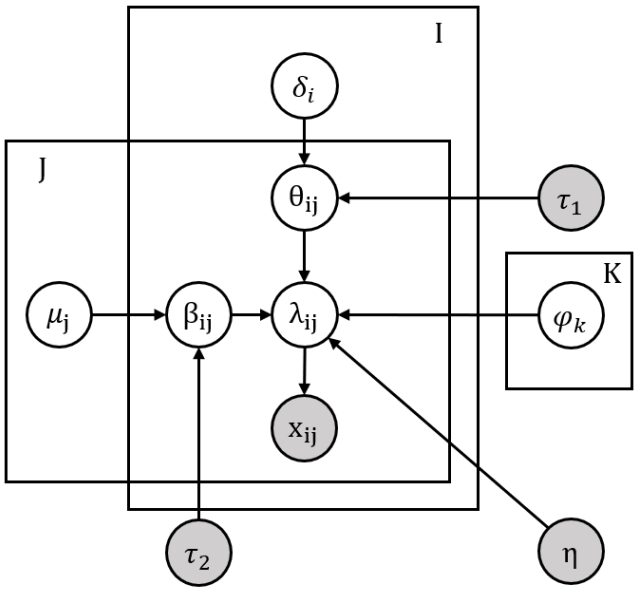
$$\beta_{ij} \sim Truncated\ Normal(\mu_j, \tau_2)$$

For each droplet i, we specify a parameter δ_i as the mean for the ambient contamination level θ_{ij}. θ_{ij} follows a truncated normal distribution bounded by (0,1).

$$\theta_{ij} \sim Truncated\ Normal(\delta_i, \tau_1)$$

Both τ_1 and τ_2 are prior parameters specified a priori, and they control the belief of counts coming from contamination profiles deviating from the mean against the native profile.

The plate diagram for the model is shown below.



Generation Process

Given cell type $\boldsymbol{k}$, library size $\boldsymbol{L}$ for droplets, hyperparameters τ_1, τ_2:

1. For each cell type k, generate native rate ϕ_k
2. For each droplet, draw δ_i
3. For each ADT, draw μ_j
4. For each ADT in each droplet, draw:
   (a) $\theta_{ij} \sim TruncatedNormal(\delta_i, \tau_1)$
   (b) $\beta_{ij} \sim TruncatedNormal(\mu_j, \tau_2)$
   (c) $x_{ij} \sim Poisson(\lambda_{ij})$
   where $\lambda_{ij} = L_i * (1-\beta_{ij}) * \theta_{ij} * \eta_j + L_i * (1-\beta_{ij}) * (1-\theta_{ij}) * \phi_{kj} + L_i * \beta_{ij}$

Data preparation and inference method

Datasets were QC-ed before inference. Droplets with top and bottom one percent of ADT and RNA library sizes were removed. Droplets with 15% or higher mitochondrial gene counts were also removed. Cell clusters were generated using the Seurat package. The prior parameters

were set to τ_1(PBMC 10k: 2e-5, PBMC 5k: 5e-5, MALT 10k: 4e-5, Golomb: 2e-5) and τ_2(PBMC 10k: 2e-6, PBMC 5k: 5e-6, MALT 10k: 2e-6, Golomb: 2e-6). The inference was done using the variational inference implementation in Stan. θ_{ij} was initialized 1e-4 and β_{ij} was initialized 1e-2. The max number of iterations were set to 50,000.

Benchmarking with other methods

Four datasets were used in benchmarking. PBMC 10k: 7865 filtered droplets of peripheral blood mononuclear cells from a healthy donor stained with 17 Total-Seq-B antibodies. PBMC 5k: 5527 filtered droplets of peripheral blood mononuclear cells from a healthy donor stained with 31 Total-Seq-B antibodies. MALT 10k: 8412 filtered droplets of dissociated Extranodal Marginal Zone B-Cell Tumor cells stained with 17 Total-Seq-B antibodies. These three datasets were downloaded from 10x Genomics website. Golomb: 30,569 filtered droplets of brain samples of mice with antibiotics induced gut microbiota depletion. A panel of 31 antibodies was used for staining the samples of 3 young mice and 3 aged mice, and the samples were pooled together using hashtag oligo. These hashtag oligos were removed from the dataset before our analysis. The dataset was acquired from the author.

The clustering of datasets was generated using the Seurat package. Silhouette widths were calculated on datasets after CLR normalization and averaged by clusters. The positive and negative scores for each dataset were calculated by finding the percentage of droplets in the dataset that belong to the clusters and have the markers expressed more than the threshold after decontamination. The threshold was set to 1 except when evaluating dsb, we also used a high threshold of 5 (reported the results as “dsb (high threshold)” in the main manuscript). The cell clusters and markers used to calculate the scores are listed in **Supplementary Table S1**.

References

References

1. Stoeckius, M. *et al*. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865-868 (2017).
2. Nathan, A. *et al*. Multimodally profiling memory T cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nat. Immunol.* **22**, 781-793 (2021).
3. Cadot, S. *et al*. Longitudinal CITE-Seq profiling of chronic lymphocytic leukemia during ibrutinib treatment: evolution of leukemic and immune cells at relapse. *Biomarker Research* **8**, 72 (2020).
4. Wu, S. Z. *et al*. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334-1347 (2021).
5. Ben-Chetrit, N. *et al*. Integration of whole transcriptome spatial profiling with protein markers. *Nat. Biotechnol.* (2023).
6. Mimitou, E. P. *et al*. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature methods* **16**, 409-412 (2019).
7. Zheng, Y., Jun, S., Tian, Y., Florian, M. & Gottardo, R. Robust Normalization and Integration of Single-cell Protein Expression across CITE-seq Datasets. *bioRxiv*, 2022.04.29.489989 (2022).
8. Gayoso, A. *et al*. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods* **18**, 272-282 (2021).
9. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151 (2020).
10. Yang, S. *et al*. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
11. Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature Communications* **13**, 2099 (2022).
12. Sheng, C. *et al*. Probabilistic machine learning ensures accurate ambient denoising in droplet-based single-cell omics. *bioRxiv*, 2022.01.14.476312 (2022).
13. Golomb, S. M. *et al*. Multi-modal Single-Cell Analysis Reveals Brain Immune Landscape Plasticity during Aging and Gut Microbiota Dysbiosis. *Cell Reports* **33**, 108438 (2020).
14. Hao, Y. *et al*. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
15. Grieshaber-Bouyer, R. *et al*. The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments. *Nature Communications* **12**, 2856 (2021).
16. Kim, M., Lu, R. J. & Benayoun, B. A. Single-cell RNA-seq of primary bone marrow neutrophils from female and male adult mice. *Scientific Data* **9**, 442 (2022).
17. Jubel, J. M., Barbati, Z. R., Burger, C., Wirtz, D. C. & Schildberg, F. A. The Role of PD-1 in Acute and Chronic Infection. *Frontiers in immunology* **11**, 487 (2020).
18. Okazaki, T., Chikuma, S., Iwai, Y., Fagarasan, S. & Honjo, T. A rheostat for immune responses: the unique properties of PD-1 and their advantages for clinical application. *Nat. Immunol.* **14**, 1212-1218 (2013).
19. Francisco, L. M., Sage, P. T. & Sharpe, A. H. The PD-1 pathway in tolerance and autoimmunity. *Immunol. Rev.* **236**, 219-242 (2010).