Improving oligo-conjugated antibody signal in multimodal single-cell analysis

Terkild Brink Buus[1,2*], Alberto Herrera[1], Ellie Ivanova[1], Eleni Mimitou[3], Anthony Cheng[4,5], Thales Papagiannakopoulos[1], Peter Smibert[3], Niels Ødum[2], & Sergei B. Koralov[1*]

[1]Department of Pathology, New York University School of Medicine, New York, NY, USA.
[2]LEO Foundation Skin Immunology Research Center, Department of Immunology & Microbiology, University of Copenhagen, Copenhagen, Denmark.
[3]Technology Innovation Lab, New York Genome Center, New York, NY, USA.
[4]Department of Genetic and Genome Sciences, University of Connecticut School of Medicine, Farmington, CT, USA.
[5]Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA, USA.

* Correspondence to: Terkild Brink Buus (Terkild.Buus@sund.ku.dk) or Sergei B. Koralov (Sergei.Koralov@nyulangone.org)

Abstract

Simultaneous measurement of surface proteins and gene expression within single cells offers high resolution snapshots of complex cell populations. These methods rely on staining cells with oligo-conjugated antibodies analogous to staining for flow- and mass cytometry. Unlike flow- and mass cytometry, signal from oligo-conjugated antibodies is not hampered by spectral overlap or limited by the number of metal isotopes, making it a highly sensitive and scalable approach. Signal from oligo-conjugated antibodies is quantified by counting reads from high-throughput sequencing. Consequently, cost of sequencing is strictly dependent on the signal intensities and background from the pool of antibodies used in analysis. Thus, considering the “cost-of-signal” as well as optimizing “signal-to-noise”, makes titration of oligo-conjugated antibody panels more complex and even more important than for flow- and mass cytometry. In this study, we investigated the titration response of a panel of oligo-conjugated antibodies towards four variables: Antibody concentration, staining volume, cell number at staining, and tissue of origin. We find that staining with high antibody concentrations recommended by published protocols and commercial vendors cause unnecessarily high background signal and that concentrations of many antibodies can be drastically reduced without loss of biological information. Reducing staining volume only affects antibodies targeting highly abundant epitopes used at low concentrations and can be counteracted by reducing cell numbers at staining. We find that background signal from empty droplets can account for a major fraction of the total sequencing reads and is primarily derived from antibodies used at high concentrations. Together, this study provides new insight into the titration response and background signal of oligo-conjugated antibodies and offers concrete guidelines on how such panels can be improved.

Introduction

Analysis of surface proteins in multimodal single-cell genomics such as cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) is a powerful addition to conventional single-cell RNA sequencing (scRNA-seq) [1, 2]. Unlike flow- and mass cytometry, CITE-seq is not limited by spectral overlap nor availability or distinguishable isotopes, respectively. This is due to the practically unlimited number of distinct oligo barcodes and discrete sequence counting, allowing high numbers of antibodies to be included in individual experiments.

While signal acquisition in CITE-seq is different, the reagents and staining procedure is highly analogous to staining for flow cytometry. Traditional titration for flow cytometry aims to identify the conjugated antibody concentration allowing the best discrimination between the signal from positive and negative cells [3]. Multiple factors may affect antibody binding and subsequent signal including antibody concentration, total amount of antibody, as well as the level of target expression (epitope amount). Epitope amount is governed by the number of cells and the per-cell expression of the target epitope. These factors are in turn influenced by the cellular composition of the sample as well as their activation and differentiation state. Nonspecific binding is expected to increase as the total amount of antibody molecules greatly exceed the epitopes present in a sample. As such, nonspecific binding is dependent on the total number of antibody molecules rather than the antibody concentration. This makes staining volume, cell composition and cell number important parameters for optimal staining [3]. Consequently, flow cytometric optimization aims to use antibody concentrations that reaches the highest signal to noise ratio (often reached at the “saturation plateau”) in a minimal volume (and thus minimal number of antibody molecules).

Oligo-conjugated antibody signal has been shown to be highly analogous to fluorochrome-conjugated antibodies of the same clone in flow cytometry in regards to the concentration needed to reach the “saturation plateau” [4]. However, unlike flow cytometry, where antibody (fluorescence) signal intensity has no influence on analysis cost, oligo-conjugated antibody signal is analyzed by counting sequencing reads, making costs strictly dependent on signal intensity (by requiring increased sequencing “depth”). This is particularly important for methods sequencing vast numbers of cells stained with a high number of antibodies such as single cell combinatorial indexed cytometry by sequencing (SCITO-seq), where shallow sequencing is paramount for the economic feasibility of such methods [5]. Thus, while an optimal antibody concentration in flow cytometry aims to get the highest signal-to-noise ratio, oligo-conjugated antibody staining conditions should be titrated to get “sufficient” signal-to-noise at the lowest possible signal intensity. In practice, this means that concentrations of most antibodies in an optimized CITE-seq panel are not intended to reach their "saturation plateau", but should be within their "linear" concentration range (where doubling the antibody concentration leads to twice the signal). Such concentrations are much more sensitive to the number of available epitopes (i.e. cell number and cell composition) than an optimized flow cytometry panel. Unlike flow cytometry where the major source of background is autofluorescence and nonspecific binding of the antibodies, a major source of background signal for oligo-conjugated antibodies appears to be free-floating antibodies in the cell suspension [6]. In droplet-based single-cell sequencing methods, these free-floating antibodies will be distributed between cell-containing and empty droplets. As signal from empty droplets can only be distinguished from signal from cell-containing droplets after sequencing and due to the much higher number of empty than cell-containing droplets, background signal can make up a considerable fraction of the sequenced reads, and thus sequencing costs.

In this study, we present a limited but practically applicable titration of four variables in a 5'-CITE-seq panel of 52 antibodies: 1. Antibody concentration (four-fold dilution response), 2. Staining volume (50 µl vs 25 µl), 3. Cell count ($1x10^6$ vs. $0.2x10^6$) and 4. Tissue of origin: peripheral blood mononuclear cells (PBMCs) from healthy donor versus immune cell compartment from a lung tumor sample. We find that oligo-conjugated antibodies show high background and little response to titration when used above 2.5 µg/mL and that most antibodies appear to reach their saturation plateau at concentrations between 0.62 and 2.5 µg/mL. Many antibodies can be further diluted despite being at their "linear" concentration range without affecting the identification of epitope-positive cells. Reducing staining volume has a minor effect on signal and only impacts signal from antibodies used at low concentrations targeting highly expressed epitopes; this effect is counteracted by reducing the number of cells present during staining. We find that background signal in empty droplets can constitute a major fraction of the total sequencing reads and is skewed towards antibodies used at high concentrations targeting epitopes present in low amounts. Finally, we compare three pipelines (Cell Ranger, CITE-Seq-Count and kallisto-bustools KITE) for "counting" antibody-derived tags (ADTs). We find that while all three pipelines yielded comparable read assignment, they showed drastic differences in their runtime and in their default unique molecular identifiers (UMI) correction approach.

Results

Four-fold antibody dilution in PBMC and lung tumor immune cells

A panel of 52 oligo-conjugated antibodies was allocated into several groups of starting concentrations based on previous experience with each antibody or targeted epitope abundance by CITE-seq, flow cytometry and vendor recommendations (ranging from 0.05 to 10 µg/mL; Supplementary Table 1). We stained two samples of either 10^6 PBMCs or $5x10^5$ lung tumor leukocytes in 50µl of antibody mix (Dilution Factor 1; DF1). To determine how the signal of each marker changed by dilution across the two tissues, we stained the same number of cells in the same volume with a four-times diluted antibody mixture (DF4).

Single-cell gene expression was assessed by shallow sequencing (5,000 reads per cell) to assign cells into major cell types (Fig. 1A) and transcriptional clusters (Fig. 1B). Leukocytes from lung tumor samples exhibited distinct transcriptional profiles within each cell type but showed overall good co-clustering with similar cell types (Fig. 1C). To allow direct comparison of unique molecular identifier (UMI) counts from the different samples, we reduced the number of cells included in analysis of each sample to contain the same number of cells from each transcriptional cluster. By only using the gene expression modality for cluster assignment, we can directly compare antibody-derived tag (ADT) UMI counts at different staining conditions within transcriptional sub-clusters without risk of having differences in ADT signal interfere with cluster assignment.

Comparing the total ADT UMI counts from each condition, we saw fewer reads from samples stained with DF4 as compared with DF1, at 77 % sequencing saturation (Fig. 1D). However, the reduction in UMI counts from DF1 to DF4 by 38% (761,350 to 474,404) and 51% (1,121,940 to 548,393) in PBMC and Lung, respectively, was markedly less than the four-fold difference (75% reduction) in antibody concentrations used in staining. It is worth noting, that 4/52 antibodies used at the highest concentration (10µg/mL) accounted for more than 20% of the total UMI counts irrespective of tissues and dilution factors and without showing any clearly positive populations (Fig. 1D, E; "gating thresholds" shown in Suppl. Fig. S1). Indeed, we found that the majority of antibodies used in concentrations at or above 2.5 µg/mL showed minimal response to four-fold titration, both in terms of total UMI counts (Fig. 1F) as well as UMI counts at the 90th quantile of the cluster with the highest overall expression level (Fig. 1G; expressing clusters identified in Fig. 1E), reflecting the response within the positive population where such could be identified. In contrast, antibodies used in concentrations at or below 0.62 µg/mL all showed close to linear response to four-fold dilution (shown as a reduction around two "logs" on a log2 scale; Fig. 1F, G). This indicates that the signal for many antibodies reaches saturation in the range between 0.62 and 2.5 µg/mL, and that higher concentrations are likely to only increase the background signal.

In the present panel, the response to four-fold dilution can be divided into five categories (Fig. 2 and Suppl. Fig. S2), that warrant different considerations in the choice of whether to reduce concentration or not. For category A (Fig 2A) reducing concentration is always the right choice. For the other categories (Fig. 2B-E), the choice of whether to reduce concentration or not, depends on the balance between the need for signal and the economic cost of signal (see Suppl. Table S2).

Reducing staining volume primarily affects highly expressed markers

To investigate the effect on ADT signal caused by further reducing the staining volume, we included PBMC samples stained with the same concentration of antibodies in 50 µl or 25 µl (effectively using half the amount of antibodies at twice the cell density). In both samples, we used the DF4 panel on 10^6 cells to assess the “worst-case scenario” of the reduction, as the amount of epitopes in this setting are likely to be competing for antibodies that are not in excess. Despite having many antibodies responding linearly to concentration reduction (Fig. 1), we found much less response to reduced staining volume, both in regard to total number of UMIs (9% reduced; 469541 to 428680) and on a marker by marker basis (Fig. 3A-C). As expected, antibodies used in low concentrations (0.0125 to 0.025 µg/mL) targeting highly abundant epitopes were most severely affected by the reduced staining volume (such as CD31, CD44 and CD45; Fig. 3D, E and Suppl. Fig. S3), whereas antibodies targeting less abundant epitopes were largely unaffected (such as CD8 and CD19; Fig 3F).

Reducing cell number during staining increases signal for antibodies at low concentration

To determine if the limited effect of reduced staining volume on ADT signal could be counteracted by simultaneously reducing the number of cells at the time of staining (effectively reducing the total amount of epitopes), we analyzed two PBMC samples with either $1x10^6$ or $0.2x10^6$ cells stained with the same concentration of antibodies (DF4) in 25µl. Similar to reducing staining volume, the majority of the included antibodies were largely unchanged by lowering the cell density at staining, as reflected by only a minor 8% increase in detected UMIs (from 428,680 to 462,916), and also reflected by the analogous distribution of individual markers (Fig. 4A-C). Encouragingly, reducing the cell number at staining increased the signal from the antibodies used at low concentration and targeting highly expressed epitopes (Fig. 3D, E and Suppl. Fig. S4), thus largely mitigating the loss of signal observed when the staining volume was reduced from 50 to 25 µl (Fig. 3B-D and Suppl. Fig. S4). Interestingly, despite reducing the cell density at staining 5-fold (from 40 to 8x10^6 cells/mL) the resulting signal did largely not supersede that of the sample stained in 50 µl with an intermediate cell density of 20x10^6 cells/mL (Suppl. Fig. S5).

Background signal from oligo-conjugated antibodies is dependent on antibody concentration and abundances of epitopes

Free-floating antibodies in the solution has been shown to be one of the major contributors to background signal for ADT [6]. Similar to cell-free RNA, background ADT signal can be assayed from empty droplets. To determine the background signal of the different antibodies in our panel, we split the captured barcodes into cell-containing and empty droplets based on the inflection point of the barcode-rank plot for the gene expression UMI counts (Suppl. Fig. S6). Despite being a “super-loaded” 10X Chromium run targeting 20,000 cells, the number of empty droplets vastly outnumber the cell-containing droplets. Consequently, several antibodies exhibited more cumulated UMIs within empty droplets than within cell-containing droplets (Fig. 5A). This was particularly prevalent within antibodies used at concentration at or above 2.5 µg/mL, thus drastically skewing the frequency of these antibodies within the empty droplets as compared with cell-containing droplets (Fig. 5A, B). Conversely, antibodies targeting highly abundant epitopes were enriched within cell-containing droplets, irrespective of their staining concentration (such as CD44 and CD107a, HLA-ABC, HLA-DR; Fig. 5C). Enrichment of antibodies targeting abundant epitopes (such as CD3, CD4, CD8 and CD45RA) within the cell-containing droplets despite high numbers of UMIs within empty droplets was also observed within three publicly available datasets using an identical 17 antibody panel on 1,000 or 10,000 cells using two different capture approaches (3’- and 5’ capture; Suppl. Fig. S7). We found that ADT signal in empty droplets (i.e. background) was highly correlated with the UMI cutoff for detection (Fig. 5D; Suppl. Fig. S8). Markers with low background generally showed low UMI cutoff and exhibited high dynamic range allowing identification of multiple levels of expression (as seen for CD4 and CD19; Fig. 5D, E). In contrast, markers with high background showed high UMI cutoff regardless of whether they exhibited cell type-specific signal (such as CD86 and CD279; Fig. 5F) or whether their positive signal was absent or obscured by the high background (such as TCRγδ; Fig. 5G).

ADT counting methods have minor influence on the resulting count matrix

Alignment and UMI counting tools can affect the resulting expression matrix for gene expression modalities in scRNA-seq [9]. Furthermore, different tools have marked differences in their runtime and computation requirements [9, 10]. To determine if this is also the case for UMI counting methods for ADT, we compared the runtime and assignment of sequencing reads to a unique CITE-seq antibody, cell barcode and UMI combination using three methods: CITE-seq-Count, Cell Ranger (running in featureOnly mode; 10X Genomics) and Kallisto-bustools (using the KITE pipeline) [10, 11]. To make the results more broadly applicable, in addition to the 52 antibody ADT library (ADT), we also included the 6 antibody cell hashing library used to demultiplex the samples (HTO) as well as three publicly available 17 antibody panel datasets where raw data is available (from the 10X Genomics website).

We found that the three methods had almost identical read to antibody assignment rates but showed drastic differences in their runtime (Fig. 6A, B). Further, while the total number of assigned reads were similar, Cell Ranger and Kallisto-bustools reported higher numbers of UMIs as compared with CITE-seq Counter run using default UMI correction parameters. This difference was due to differences in default behavior in regard to whether “errors” in UMIs are corrected during UMI collapsing or not by the different methods (Fig. 6C, D). This difference was particularly evident within datasets stained with few oligo-conjugated antibodies (such as the HTO dataset) and within cells containing the highest amount of ADT counts, where similar UMIs are more likely to be assigned to the same antibody/cell barcode combination (Fig. 6C, D and Suppl. Fig. S9).

Discussion

In this study, we show that titration of oligo-conjugated antibodies for multimodal single-cell analysis can improve the sensitivity, reduce sequencing requirements and spare costs, and that such optimizations go beyond (and even against) the need to reach the “saturation plateau”. We show that for a representative panel of 52 antibodies, most antibodies used in concentrations at or above 2.5 µg/mL show high background signal and minimal loss in sensitivity upon a four-fold reduction in concentration. Antibodies used at concentrations between 0.625 and 2.5 µg/mL show limited (non-linear) response whereas most antibodies used at concentrations below 0.625 µg/mL show linear or close to linear response. It should be noted, that these estimates may be inherently biased given that the starting concentrations were based on our prior experience with the individual antibody clones and our assumptions regarding abundance of targeted epitopes. This has favored using higher concentrations for antibodies known to have low “performance” and for antibodies with unknown performance. Nonetheless, for antibodies with unknown performance, our results highlight the benefits of conducting titration experiments or initially using the antibodies at concentrations in the 0.625 to 2.5 µg/mL range rather than the 5 to 10 µg/mL range recommended by published antibody staining protocols and by commercial vendors. This is particularly important when adding new antibodies to existing panels, where antibodies added in a high concentration may account for a drastically disproportionate usage of the total sequencing reads without providing any biological information (as seen for CD86, CD152, CD183, CD197 and TCRgd in the DF1 panel). Our results also show that concentrations of antibodies targeting highly expressed epitopes can be further reduced without affecting resolution of positive and negative cells, even when these antibodies are already used within their linear concentration range. By reducing the concentration of these antibodies, the allocation of reads to each antibody becomes more balanced between more and less abundant epitopes allowing the overall sequencing depth to be reduced and maximizing the yield of a sequencing run.

Reducing staining volume for 10^6 PBMCs from 50 µl to 25 µl only showed minor effect on signal and this minimal impact was primarily observed for antibodies used at very low concentrations (0.0125 to 0.025 µg/mL) targeting highly expressed epitopes (such as CD31, CD44 and CD45). This effect was readily counteracted by concomitantly reducing the number of cells at staining to 0.2×10^6 PBMCs in 25 µl. In flow cytometry, while the binding of antibody is strictly dependent on the concentration, background signal is dependent on the ratio between the total amounts of antibody and epitopes [3]. Consequently, background can be reduced by increasing the number of cells (increasing the amount of epitope) or decreasing staining volume (effectively reducing the amount of antibody without changing its concentration). For antibodies optimized to reach their “saturation plateau” (common in flow cytometry), both of these approaches can be applied without changing the true signal. In contrast, for oligo-conjugated antibodies used in sequencing based single-cell approaches, operating in the “linear” range, signal from highly abundant epitopes stained with low concentration of antibody will be affected. In such cases, the cells can be stained in multiple steps adjusting the staining volume while keeping the concentration the same – i.e. staining in a smaller volume for antibodies with high background and subsequently staining antibodies at low concentration in a higher volume. In this regard, when multiplexing samples, pre-staining each sample with hashtags and pooling prior to staining with additional CITE-seq antibodies may provide multiple advantages: 1. All samples are stained at the same time with the exact same antibody mixture – making cross-sample comparison more accurate. 2. By having more cells in a smaller total volume, less total antibody is used in the presence of more epitopes conceivably reducing the background signal. 3. Samples where cell number at staining is a limiting factor, such as small tissue biopsies, will be exposed to the same local concentrations of antibody as more abundant samples (such as PBMCs) removing potential differences between samples by antibodies being "sponged" by differences in overall epitope abundance. However, this approach is only available when all samples are similarly affected by the staining procedure and can tolerate the additional washes needed (after both hashing and CITE-seq staining).

Empty droplets have been shown to be useful for determining the background signal of CITE-seq [6]. This suggests that the major source of background signal for ADT libraries can be attributed to free-floating antibodies (or oligos) in the solution, rather than unspecific antibody binding to cell surfaces. In the present study, the samples were multiplexed by hashing antibodies and pooled after oligo-conjugated antibody staining and then run in the same 10X Chromium lane. This effectively obscures the contribution of each sample to the total amount of free-floating antibodies in the final cell suspension which is conceivably skewed towards the samples stained in high volume with the highest concentration of antibodies – as these samples contain the highest total amount of antibody. Consequently, as free-floating antibodies are the major source of background, this would explain why we do not observe reduced background in the cell stained at the lowest concentrations (i.e. dilution factor 4). As such, for markers with no specific signal due to high background (such as CD183, CD197 and TCRgd), the titration responses may be underestimated due to specific signal being lost within the high background. This also means that for markers with high background signal, our proposed reductions in concentrations are conservative, as we would expect to see decreased background in samples stained with reduced amount of antibodies. In droplet-based single-cell analyses, background signal is not only diminishing the sensitivity and resolution of true signals, it is also a major contributor to sequencing cost of ADT libraries. Due to empty droplets vastly outnumbering cell-containing droplets, we found that ADT signal from empty droplets can easily account for 20-50% of the total sequencing reads. The number of antibodies used in CITE-seq-related platforms is only expected to expand. Additionally, the numbers cells included in each experiment is continuously being increased (as seen for methods such as SCITO-seq [5]). As such, reducing background signal from oligo-conjugated antibodies should be a priority. The source of the free-floating antibodies is not completely understood. Observations from this study suggest that antibodies used at high concentration targeting absent or sparse epitopes are highly enriched within the empty droplets as compared with the cell-containing droplets indicates that residual antibody from the staining step is a major contributor, despite several washing steps. Practically, this suggests that additional washing after cell staining would be beneficial when the number and type of cells in the samples allow it. Optimal washing is achieved by repeated washing steps while assuring that maximal residual supernatant is removed after each centrifugation and followed by gentle but complete resuspension in a large buffer volume.

We compared three pipelines for counting ADT reads and found that while all three methods resulted in similar assignment rates, they exhibited drastic differences in their runtimes. The kallisto-bustools KITE pipeline was an order of magnitude faster than the other methods, without showing any impact on the resulting count matrices. Importantly, we found that CITE-seq-Count consistently returned lower UMI counts than kallisto-bustools KITE and Cell Ranger despite assigning a similar number of reads. This difference could be attributed to differences in UMI correction, as the UMI counts were comparable when CITE-seq-Count was instructed not to correct UMIs. By default, CITE-seq-Count collapses two UMIs if their sequences have a hamming distance of 2 or less and they are assigned to the same cell and antibody barcode. In contrast to genome-wide gene expression, the limited number of targets mean that ADT sequences are much more likely to contain “similar” UMIs assigned to the same cell and antibody barcode by chance. If UMI correction is warranted, we would expect to see the corrected UMIs and consequently, the reduced UMI counts evenly distributed among all cells and antibody barcodes. In contrast, we see that the reduced UMI counts are most significant in libraries with few targets (such as the HTO library with 6 antibodies) and tend to be found within abundant antibody barcodes of cells with high overall UMI counts. So, while UMI correction may be necessary for some protocols (i.e. using sequencers with higher error rates or requiring more PCR cycles), we find that UMI correction may be detrimental for CITE-seq libraries of the sizes investigated in this study.

More and more advanced CITE-seq-related “cytometry-by-sequencing” platforms are rapidly being developed. However, while these platforms utilize different methods to assure single-cell resolution, and use different approaches to label the cells, they all use high-throughput sequencing to count signal from a variety of oligo-conjugated probes (such as antibodies, MHC-peptide multimers, B-cell receptor antigens etc.) [1, 2, 5, 12-14]. Most of the observations, results and conclusions from this study will thus also be applicable to a variety of platforms where improving oligo-conjugated probe signal is critical for ensuring their broad utility and economic feasibility.

Materials and methods

Clinical samples

Patient and control samples were collected at New York University Langone Health Medical Center in accordance with protocols approved by the New York University School of Medicine Institutional Review Board and Bellevue Facility Research Review Committee (IRB#: i15-01162 and S16-00122).

Cell isolation, cryopreservation and thawing

Peripheral blood mononuclear cells (PBMCs) were isolated from the blood of a healthy volunteer by gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare) and Sepmate-50 tubes (Stemcell). Buffy coat PBMCs were collected and washed twice with PBS 2% FBS. Lung tumor sample were cut into small pieces with a razor blade and enzymatically digested (100 U/mL Collagenase IV, Sigma-Aldrich, C5138-1G; 50 µg/mL DNase 1, Worthington, LS002138) for 35 minutes being rotated at 37°C in HEPES buffered RPMI 1640 containing 0.5% FBS. After digestion, the sample was forced through a 100 µm cell strainer to make a single-cell suspension. Single-cell suspensions from both PBMCs and lung tumor were cryopreserved in freezing medium (40% RPMI 1640, 50% FBS and 10% DMSO) and stored in liquid nitrogen. On the day of the experiments, cryopreserved samples were thawed for 1-2 minutes in a 37°C water bath, washed twice in warm PBS containing 2% FBS and re-suspended in complete media (RPMI 1640 supplemented with 10% FBS and 2mM L-Glut).

Oligo-conjugated antibody staining

We modified the published protocol for ECCITE-seq [7] to stain cells in round-bottom 96-well plates (as is common practice for flow cytometry staining in many laboratories). This allowed us to reduce staining volumes and centrifugation time analogous to staining for flow cytometry. After thawing, the intended number of cells were resuspended in half the intended staining volume of CITE-seq staining buffer (2% BSA, 0.01% Tween in PBS). To prevent antibody binding to Fc receptors, single-cell suspensions were incubated for 10 min with 1X Fc receptor block from two vendors (TruStain FcX, BioLegend and FcR blocking reagent, Miltenyi) in half the intended staining volume. During incubation, the antibody solution of 52 TotalSeqC antibodies (BioLegend; Suppl. Table 1) was washed on a pre-wet Amicon Ultra-0.5 Centrifugal Filter to remove sodium azide. The volume of the resulting antibody pool was adjusted to 2X of final concentrations and was added to the cells in half the intended staining volume. 10µg/mL of a unique hashing antibody was added to each sample and incubated for 30 min on ice. After staining, cells were washed four times in 1x150 µl and 3x200 µl CITE-seq staining buffer.

Super-loading of 10X chromium

Individually hashed samples were counted using a hemocytometer and pooled in equal ratio at high concentration. Pooled sample was strained through a 70µm cell strainer and counted again using a hemocytometer. To achieve approximately 20,000 cells after doublet removal, cell concentration was adjusted to 1314 cells/µl to achieve the target of 41,645 cells in 31.7µl for super-loading of the 10X Chromium Chip A. Gene expression and antibody tag libraries were constructed using reagents, primers and protocol from the published ECCITE-seq protocol [7].

Alignment and counting of single-cell sequencing libraries

The multiplexed gene expression library was aligned using kallisto (v0.46)-bustools (v0.39.0) [8, 9]. Given the polyA selection inherent in the 10X genomics protocol, reads were aligned against a reference transcriptome based on the GTF file included in the Cell Ranger software (refdata-cellranger-GRCh38-3.0.0/genes/genes.gtf; 10X Genomics) that does not include as many non-polyA transcripts as the human transcriptome included by kallisto-bustools by default. Antibody-derived tag (ADT) and hashtag-oligo (HTO) libraries were counted using the *kallisto indexing and tag extraction* (KITE) workflow (https://github.com/pachterlab/kite).

Single-cell demultiplexing, preprocessing and down-sampling

To allow detection of UMI counts within non-cell-containing droplets, unfiltered count matrices from each modality was loaded into a ‘Seurat’ (v3.1.4) object [8]. Samples were demultiplexed by their unique hashtag oligos (HTO) using the Seurat function ‘MULTIseqDemux’. This allowed the removal of all cross-sample doublets. Due to the shallow sequencing of the mRNA library, expression at least 60 genes and a percent mitochondrial reads below 15 % were used to remove barcodes from non-viable cells or debris. Intra-sample doublets were removed using the ‘scDblFinder’ (v1.1.8) R package. UMI counts from antibody derived tags (ADTs) were normalized using default configuration of the DSB (v0.1.0) R package with ADT signal from HTO-negative droplets used as empty drop matrix and using included isotype controls [6]. Gene expression was preprocessed using the default Seurat v3 pipeline and fine-grained clusters were identified using the ‘FindClusters’ function with a resolution of 1.2. Clusters were divided into major cell types by their distinct expression of lineage markers either within the mRNA or (ADT) modality. To allow direct comparison of UMI counts across conditions, each condition was down-sampled by tissue of origin to include the same number of cells within each fine-grained cluster (resulting in 1,777 cells from each PBMC sample and 1,681 cells from each Lung tumor sample).

Comparing ADT signal from cell-containing and empty droplets

For comparison of UMI counts within cell-containing and non-cell-containing (empty) droplets for the present dataset and the 10X Genomics datasets, we divided the unfiltered count matrices by the inflection point in their ranked per cell UMI sum from the mRNA library. Barcodes above the inflection point were then used to extract UMI counts within cell-containing droplets from each antibody oligo modality. All UMIs that were not included in cell-containing droplets were considered from empty droplets.

Comparison of ADT counting methods

Assignment of reads to antibody tags and barcodes were compared between three algorithms: kallisto-bustools KITE workflow (see above), CITE-seq-Count (v1.4.3) allowing a hamming distance of 1 (--max-error 1) between antibody tag and read sequence with or without UMI correction (--no_umi_correction) and Cell Ranger (v3.1.0) in feature-only mode with minimal auxiliary processing (--nosecondary --nopreflight --disable-ui). Runtimes were recorded using the ‘time’ UNIX command and included all parts of the workflows (including conversion of feature reference from the Cell Ranger format and indexing). All pipelines were run on the NYU Medical Center computing cluster on a node with 16 cores and 64GB memory.

Data and code availability

All code and commands used to process the data and to generate all plots and figures are available at GitHub: https://github.com/Terkild/CITE-seq_optimization

UMI count matrices from the optimization experiment have been deposited at FigShare with DOI: https://doi.org/10.6084/m9.figshare.c.5018987. The feature barcode 3’ and 5’ VDJ 10X datasets are available from the 10X Genomics website.

Conflict of interest

PS is co-inventor of a patent related to the single cell technology utilized in this study

Acknowledgements

Work in Dr. Koralov's laboratory was supported by the NIH R01 grant (HL-125816), LEO Foundation Grant (LF-OC-20-000351), NYU Cancer Center Pilot Grant (P30CA016087), the Judith and Stewart Colton Center for Autoimmunity Pilot grant, and a grant from the Drs. Martin and Dorothy Spatz Foundation. TBB and NØ are supported by the Danish Cancer Society (Kræftens Bekæmpelse), the Danish Council for Independent Research (Danmarks Frie Forskningsfond) and the LEO Foundation. We thank NYU Genome Technology Center for technical assistance and support and acknowledge NYU Center for Biospecimen Research & Development and NYU Perlmutter Cancer Center for their support in acquiring patient biospecimens.

Contributions:

TBB, AH, EI and SBK initiated the project and were responsible for experiment design with input from EPM, NØ, TP and PS. TBB and AH performed the experiments. TBB performed the bioinformatic data analysis with input from AC, EPM, PS and SBK. TBB, AH, EI and SBK were responsible for the writing, with all authors providing input on the manuscript.

References

1. Stoeckius, M., et al., *Simultaneous epitope and transcriptome measurement in single cells.* Nat Methods, 2017. **14**(9): p. 865-868.
2. Peterson, V.M., et al., *Multiplexed quantification of proteins and transcripts in single cells.* Nat Biotechnol, 2017. **35**(10): p. 936-939.
3. Hulspas, R., *Titration of fluorochrome-conjugated antibodies for labeling cell surface markers on live cells.* Curr Protoc Cytom, 2010. **Chapter 6**: p. Unit 6 29.
4. Stoeckius, M., et al., *Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics.* Genome Biol, 2018. **19**(1): p. 224.
5. Hwang, B., et al., *SCITO-seq: single-cell combinatorial indexed cytometry sequencing.* bioRxiv, 2020: p. 2020.03.27.012633.
6. Mulè, M.P., A.J. Martins, and J.S. Tsang, *Normalizing and denoising protein expression data from droplet-based single cell profiling.* bioRxiv, 2020: p. 2020.02.24.963603.
7. Mimitou, E.P., et al., *Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells.* Nat Methods, 2019. **16**(5): p. 409-412.
8. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
9. Du, Y., et al., *Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment.* G3 (Bethesda), 2020.
10. Melsted, P., et al., *Modular and efficient pre-processing of single-cell RNA-seq.* bioRxiv, 2019: p. 673285.
11. Melsted, P., V. Ntranos, and L. Pachter, *The barcode, UMI, set format and BUStools.* Bioinformatics, 2019. **35**(21): p. 4472-4473.
12. Setliff, I., et al., *High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity.* Cell, 2019. **179**(7): p. 1636-1646 e15.
13. O'Huallachain, M., et al., *Ultra-high throughput single-cell analysis of proteins and RNAs by split-pool synthesis.* Commun Biol, 2020. **3**(1): p. 213.
14. Overall, S.A., et al., *High throughput pMHC-I tetramer library production using chaperone-mediated peptide exchange.* Nat Commun, 2020. **11**(1): p. 1909.

Figures and Figure Legends

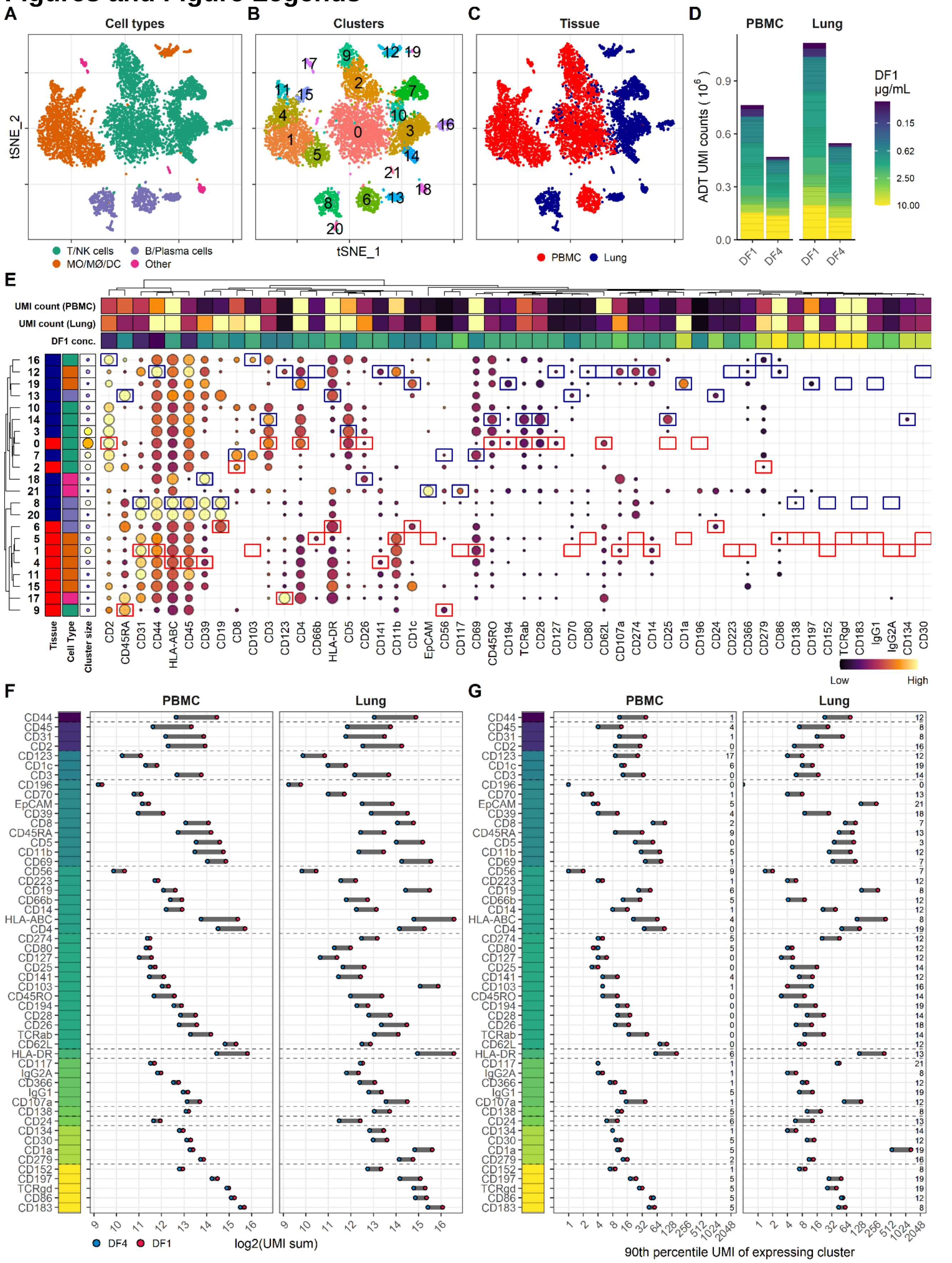

Figure 1: Four-fold antibody dilution response in PBMC and lung tumor immune cells.

A-C. Single cells from all samples and conditions were clustered and visualized according to their gene expression and colored by (**A**) overall cell type, (**B**) transcription-based cluster, and (**C**) tissue of origin. **D**. Summarized UMI counts within cell-containing droplets segmented by the individual antibodies between dilution factor 1 (DF1) and DF4 in PBMC and Lung samples. Antibody segments are colored by their concentration at DF1. **E**. Heatmap of normalized antibody-derived tag (ADT) signal within each transcription-based cluster identified in B. Visualized by frequency of positive cells (circle size) and colored by the median ADT signal within the positive fraction (i.e. signal from a marker that is highly expressed by all cells in a cluster will have the biggest circle and be colored yellow). Red and blue colored boxes denote the clusters chosen for evaluating titration response within blood and lung samples, respectively. **F, G**. Change in ADT signal for each antibody by four-fold dilution. Individual antibodies are colored by their concentration at DF1 and quantified by (**F**) sum of UMIs within cell-containing droplets assigned to each antibody and (**G**) 90th percentile UMI count within expressing cell cluster identified in E and annotated by numbers to the right. MO/MØ/DC: monocyte, macrophage or dendritic cell.

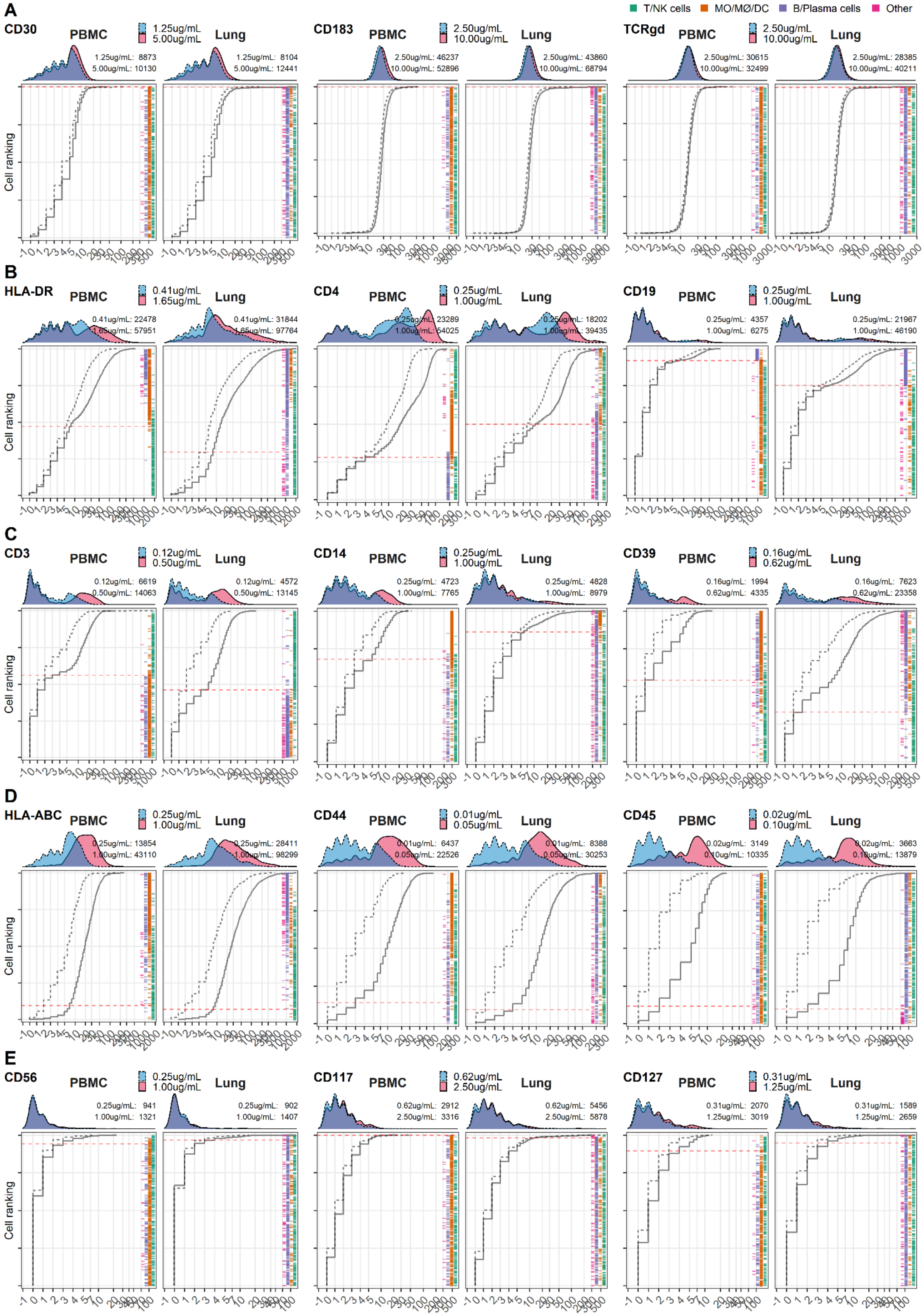
A
🟩 T/NK cells
🟧 MO/MØ/DC
🟪 B/Plasma cells
🟥 Other
CD30
PBMC
Lung
1.25ug/mL
5.00ug/mL
1.25ug/mL: 8873
5.00ug/mL: 10130
1.25ug/mL: 8104
5.00ug/mL: 12441
CD183
PBMC
Lung
2.50ug/mL
10.00ug/mL
2.50ug/mL: 46237
10.00ug/mL: 52896
2.50ug/mL: 43860
10.00ug/mL: 68794
TCRgd
PBMC
Lung
2.50ug/mL
10.00ug/mL
2.50ug/mL: 30615
10.00ug/mL: 32499
2.50ug/mL: 28385
10.00ug/mL: 40211
Cell ranking
B
HLA-DR
PBMC
Lung
0.41ug/mL
1.65ug/mL
0.41ug/mL: 22478
1.65ug/mL: 57951
0.41ug/mL: 31844
1.65ug/mL: 97764
CD4
PBMC
Lung
0.25ug/mL
1.00ug/mL
0.25ug/mL: 23289
1.00ug/mL: 54025
0.25ug/mL: 18202
1.00ug/mL: 39435
CD19
PBMC
Lung
0.25ug/mL
1.00ug/mL
0.25ug/mL: 4357
1.00ug/mL: 6275
0.25ug/mL: 21967
1.00ug/mL: 46190
Cell ranking
C
CD3
PBMC
Lung
0.12ug/mL
0.50ug/mL
0.12ug/mL: 6619
0.50ug/mL: 14063
0.12ug/mL: 4572
0.50ug/mL: 13145
CD14
PBMC
Lung
0.25ug/mL
1.00ug/mL
0.25ug/mL: 4723
1.00ug/mL: 7765
0.25ug/mL: 4828
1.00ug/mL: 8979
CD39
PBMC
Lung
0.16ug/mL
0.62ug/mL
0.16ug/mL: 1994
0.62ug/mL: 4335
0.16ug/mL: 7623
0.62ug/mL: 23358
Cell ranking
D
HLA-ABC
PBMC
Lung
0.25ug/mL
1.00ug/mL
0.25ug/mL: 13854
1.00ug/mL: 43110
0.25ug/mL: 28411
1.00ug/mL: 98299
CD44
PBMC
Lung
0.01ug/mL
0.05ug/mL
0.01ug/mL: 6437
0.05ug/mL: 22526
0.01ug/mL: 8388
0.05ug/mL: 30253
CD45
PBMC
Lung
0.02ug/mL
0.10ug/mL
0.02ug/mL: 3149
0.10ug/mL: 10335
0.02ug/mL: 3663
0.10ug/mL: 13879
Cell ranking
E
CD56
PBMC
Lung
0.25ug/mL
1.00ug/mL
0.25ug/mL: 941
1.00ug/mL: 1321
0.25ug/mL: 902
1.00ug/mL: 1407
CD117
PBMC
Lung
0.62ug/mL
2.50ug/mL
0.62ug/mL: 2912
2.50ug/mL: 3316
0.62ug/mL: 5456
2.50ug/mL: 5878
CD127
PBMC
Lung
0.31ug/mL
1.25ug/mL
0.31ug/mL: 2070
1.25ug/mL: 3019
0.31ug/mL: 1589
1.25ug/mL: 2659
Cell ranking

Figure 2: Four-fold antibody dilution response is dependent on epitope abundance.
"Titration plots" (Marker UMI count vs. Normalized cell rank) for response to reduction in antibody concentration from dilution factor 1 (DF1; solid line) to DF4 (dashed line). Histogram depicts distribution of UMIs at each condition colored by dilution factor (and annotated with concentration). Numbers in histograms denote total UMI count within cell-containing droplets at each antibody concentration within PBMC (left) and Lung (right). "Barcodes" to the right depicts cell type occurrence at the corresponding rank to visualize cell specificity of the antibody. Horizontal red lines depict cell rank cutoff for "positive" cells. Antibody response to four-fold dilution can be divided into five categories exemplified in A-E. **A**. Antibodies where the positive signal is obscured within the background signal (Category A). **B**. Antibodies that respond by a reduction in signal but without hampering the ability to distinguish positive and negative fractions (Category B). These antibodies also show strict cell type specificity (i.e. CD4 is highly expressed in T cells and intermediately expressed in MO/MØ/DC whereas CD19 is only expressed with in B/Plasma cells as shown in the barcode plot) **C**. Antibodies that respond by a reduction in both signal and changes the ability to distinguish positive from negative fractions (Category C). **D**. Antibodies targeting ubiquitously expressed markers (Category D). **E**. Antibodies that do not show a convincing positive population due to either lack of epitopes (no positive cells in either tissue) or lack of antibody binding (non-functional antibody) (Category E). Titration plots for all markers can be found Suppl. Fig. 2.

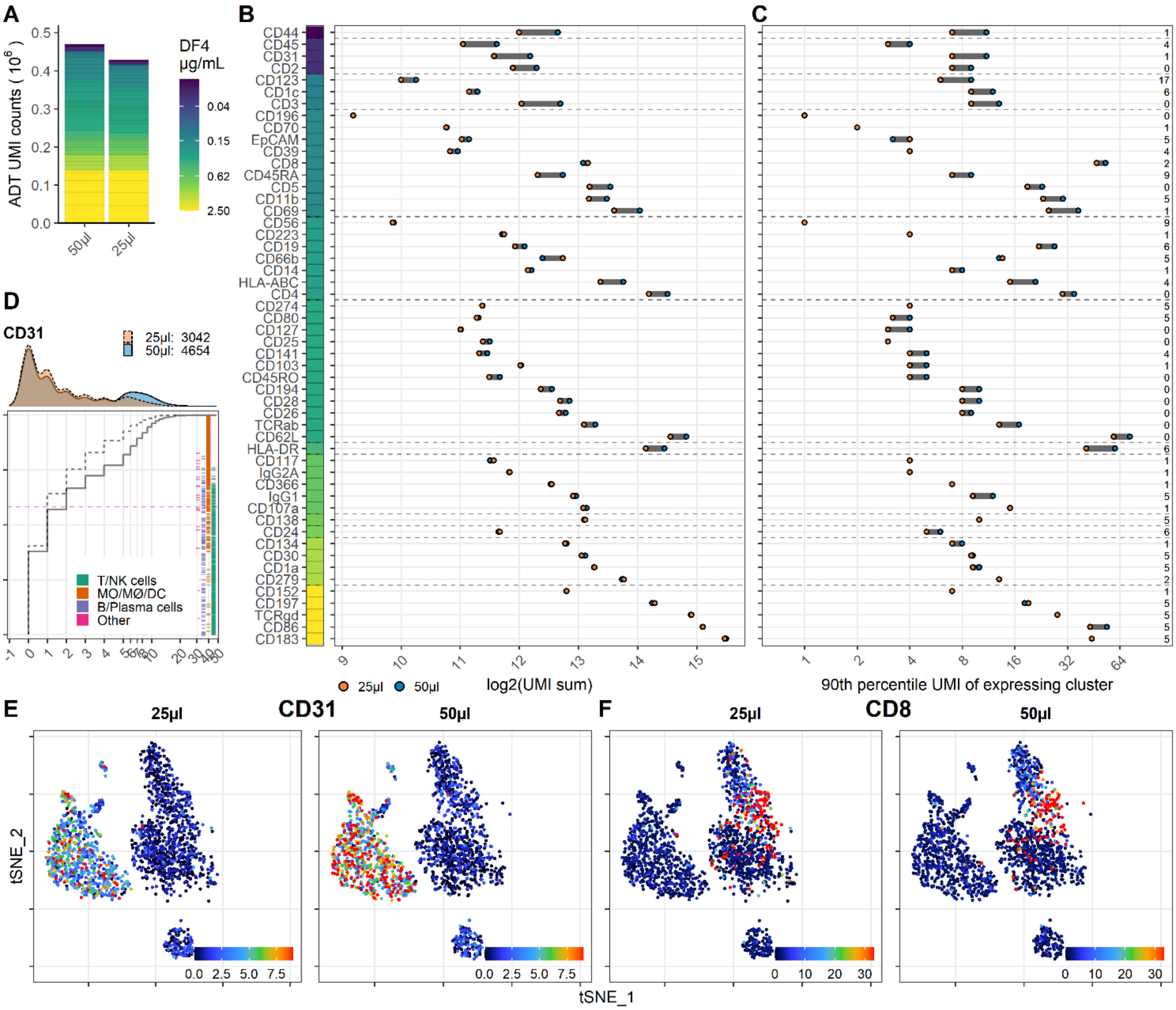

Figure 3: Reducing staining volume primarily affects highly expressed markers. Comparison of PBMC samples stained in 50µl (same sample as DF4 in Fig. 1) or 25µl volume at dilution factor 4. **A**. Summarized UMI counts within cell-containing droplets segmented by the individual antibodies colored by their concentration. **B, C**. Change in ADT signal for each antibody by reducing staining volume from 50 to 25 µl. Individual antibodies are colored by their concentration. Quantified by (**B**) sum of UMIs within cell-containing droplets assigned to each antibody and (**C**) 90th percentile UMI count within cell cluster with most abundant expression (the assayed cluster is annotated by numbers inside the). **D**. "Titration plot" (Marker UMI count vs. Normalized cell rank) for CD31 signal response to reduction in staining volume from 50 µl (solid line) to 25 µl (dashed line). Histogram depicts distribution of UMIs at each condition. "Barcode" to the right depict cell type occurrence at the corresponding rank to visualize cell specificity of the antibody. Numbers in legend denote total UMI count assigned to CD31 within cell-containing droplets from each sample. **E**, **F**. Non-normalized UMI counts visualized on tSNE of an affected (CD31; **E**) or an unaffected (CD8; **F**) marker by the reduction in cell density. Titration plots for all markers can be found Suppl. Fig. 3.

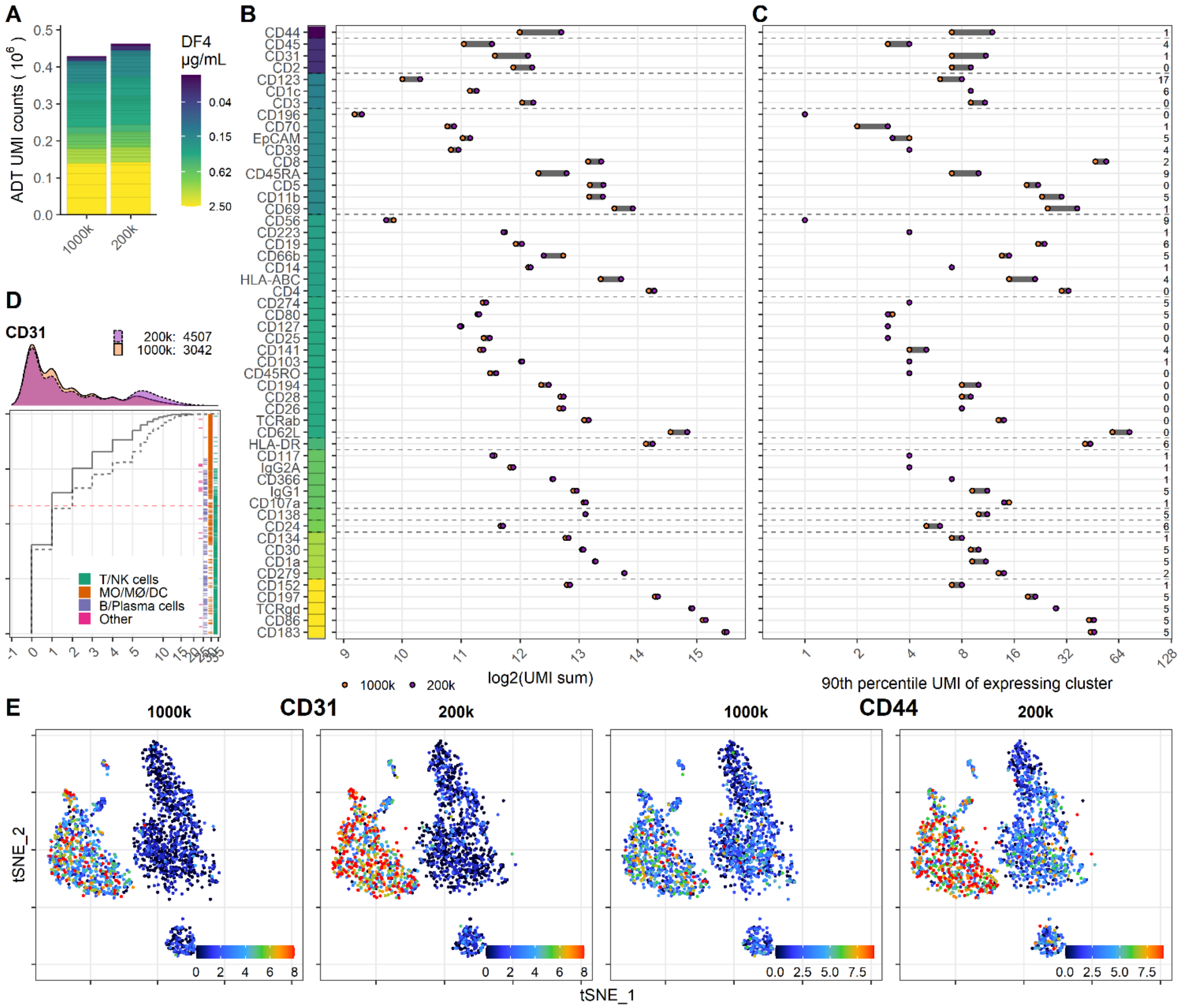

Figure 4: Reducing cell number during staining increases signal for antibodies at low concentration.

Comparison of PBMC samples stained in 25µl antibody staining solution at dilution factor 4 (DF4) at two cell densities: 1x10^6 (1000k; same sample as 25µl in Fig. 3) or 0.2x10^6 (200k) cells. **A**. Summarized UMI counts within cell-containing droplets segmented by the individual antibodies colored by their concentration. **B, C**. Change in ADT signal for each antibody by reducing cell numbers at staining from 1x10^6 to 0.2x10^6 cells. Individual antibodies are colored by their concentration. Quantified by (**B**) sum of UMIs within cell-containing droplets assigned to each antibody and (**C**) 90th percentile UMI count within cell cluster with most abundant expression (the assayed cluster is annotated by numbers inside the). **D**. "Titration plot" (Marker UMI count vs. Normalized cell rank) for CD31 signal response to reduction in cell numbers at staining from 1x10^6 (solid line) to 0.2x10^6 cells (dashed line). Histogram depicts distribution of UMIs at each condition. "Barcode" to the right depict cell type occurrence at the corresponding rank to visualize cell specificity of the antibody. Numbers in legend denote total UMI count assigned to CD31 within cell-containing droplets from each sample. **E**. Non-normalized UMI counts visualized on tSNE plot of CD31 and CD44 which affected by the reduction in staining volume, mitigated by a concomitant reduction in cell density. Titration plots for all markers can be found Suppl. Fig. 4.

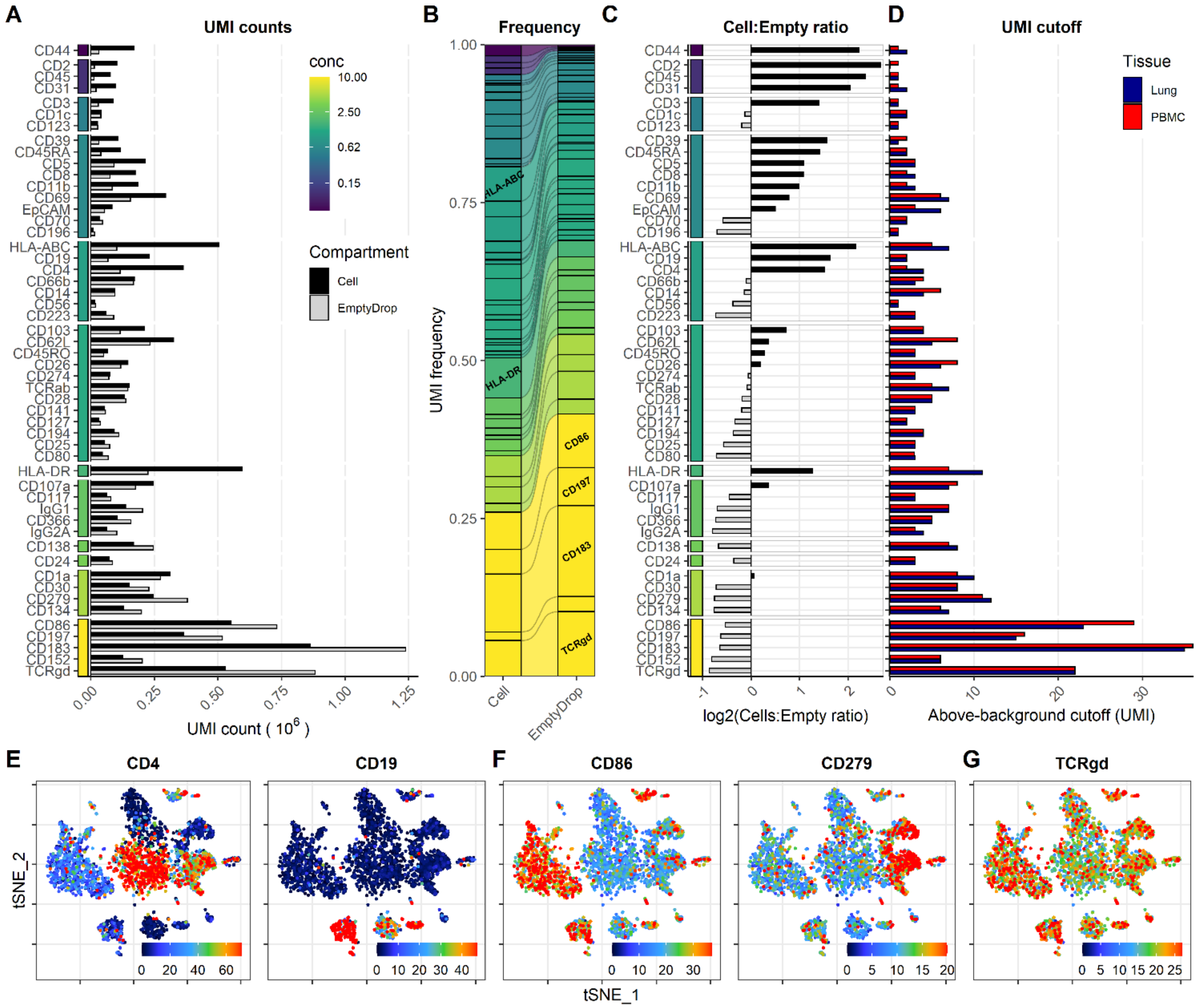

Figure 5: Background signal from oligo-conjugated antibodies are dependent on concentration and presence of epitopes.

Signal from free-floating antibodies in the cell suspension is a major source of background in droplet-based scRNA-seq and can be assayed by their signal within non-cell-containing (empty) droplets. **A-B**. Comparison of signal from each antibody within cell-containing and empty droplets (identified in Suppl. Fig. S5) by (**A**) their total UMI counts or, (**B**) their relative frequency within each compartment. Color bar denotes antibody concentration at dilution factor 1 (DF1). **C**. Ratio of UMI frequencies of each marker between cell-containing and empty droplets. Markers with black bars have greater frequency in cell-containing droplets whereas grey bars have greater frequency in empty droplets. **D**. UMI Thresholds for detection above-background for each marker within PBMC and Lung tumor samples (based on gating in Suppl. Figure S1). **E-G**. Examples of tSNE plots showing non-normalized (raw) UMI counts from cells stained at dilution factor 1 (DF1) for (**E**) markers with low background, (**F**) markers with high background that still exhibit cell type-specific signal (CD86 and CD279) and (**G**) marker where positive signal is absent or obscured by the background. To make the color scale in the tSNE plots less sensitive to extreme values, we set the upper threshold to the 90% percentile. tSNE plots for all markers can be found in Suppl. Fig. S8.

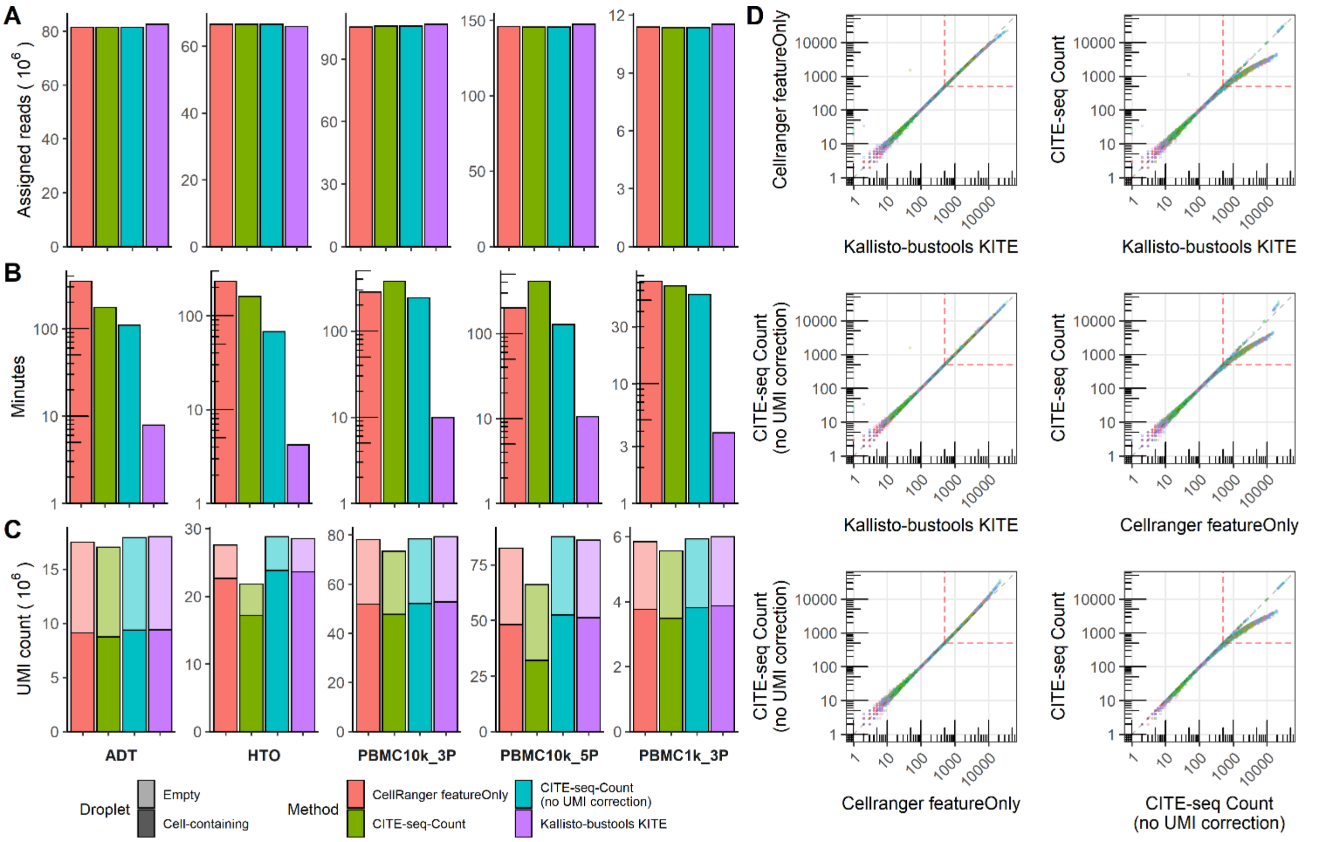

Figure 6: ADT counting methods have minor influence on count matrix but major differences in their runtime.

Comparison of ADT counting methods by (**A**) read assignment, (**B**) runtime on a high performance computing node with 16 cores and 64GB RAM, and (**C**) UMI count within cell-containing and empty droplets (identified in Suppl. Fig. S6 and S7). **D**. Pair-wise concordance of UMI count per antibody per cell-containing droplet within the HTO library. Dots are colored by antibody. Dashed red lines indicate region where CITE-seq-Count show reduced UMI counts. Pair-wise concordance plots between ADT counting methods for remaining libraries can be found in Suppl. Fig. S9.