**Large-scale simultaneous measurement of epitopes and transcriptomes in single cells.**

*Marlon Stoeckius[1,*], Christoph Hafemeister[1], William Stephenson[1], Brian Houck-Loomis[1], Harold Swerdlow[1], Rahul Satija[1,2], Peter Smibert[1]*

[1] New York Genome Center, New York, NY

[2] New York University Center for Genomics and Systems Biology, New York, NY

** Correspondence: mstoeckius@nygenome.org*

**Recent high-throughput single-cell sequencing approaches have been transformative for understanding complex cell populations, but are unable to provide additional phenotypic information, such as protein levels of well-established cell-surface markers. Using oligonucleotide-labeled antibodies, we integrate measurements of cellular proteins and transcriptomes into an efficient, sequencing-based readout of single cells. This method is compatible with existing single-cell sequencing approaches and will readily scale as the throughput of these methods increase.**

The unbiased and extremely high-throughput nature of modern scRNA-seq approaches has proved invaluable for describing heterogeneous cell populations[1-3]. Prior to the use of single cell genomics, a wealth of studies achieved detailed definitions of cellular states via cell cytometry by leveraging carefully curated panels of fluorescently labeled antibodies directed at cell surface proteins. Such approaches particularly focused on the immune and nervous systems, where well-characterized and canonical markers are reliable indicators of cellular activity and function[4]. Recent studies[5-7] have demonstrated the potential for coupling 'index-sorting' measurements with single cell transcriptomics, enabling the mapping of immunophenotypes onto transcriptomically-derived clusters. However, recently developed massively parallel approaches based on droplet microfluidics[1-3], microwells[8,9] or combinatorial indexing[10,11] do not utilize cytometry for cellular isolation, and therefore cannot couple protein information with cellular transcriptomes, representing a significant limitation for these exciting approaches.

An ideal approach for cellular phenotyping would combine the measurement of known protein markers with the detection of cellular transcripts in an unbiased manner, while maintaining the high throughput that is obtainable using current scRNA-seq technologies. Recently described attempts to simultaneously detect and/or measure transcripts and proteins in single cells employ proximity ligation or proximity extension assays in combination with digital PCR[12-15] or mass cytometry[16]. These approaches are limited in scale and/or can only profile a few genes and proteins in parallel (see Supplementary Table 1 for comparison of different technologies).
Here we describe Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq), a method that combines highly multiplexed antibody-based detection of protein markers together with unbiased transcriptome profiling for thousands of single cells in parallel. We demonstrate that the method is readily adaptable to two different high-throughput single-cell RNA sequencing applications and show by example that it can achieve a more detailed characterization of the immune system than scRNA-seq alone. CITE-seq allows analysis of cellular phenotypes at unprecedented detail in a manner that will only increase with the rapidly expanding scale of single-cell transcriptome sequencing.

We hypothesized that a DNA oligonucleotide conjugated to an antibody could be measured by sequencing as a digital readout of protein abundance. We envisioned the oligos to be compatible with oligo dT-based RNA-sequencing library preparations so that they would be captured and sequenced together with mRNAs. To this end, we designed the CITE-seq oligos with the following characteristics: a generic PCR handle for amplification and next-generation sequencing library preparation, a barcode sequence specific for each antibody, and a polyA stretch at the 3' end (Fig. 1a) designed to anneal to polyT stretches on primers used to initiate reverse transcription. We leverage the DNA-dependent DNA polymerase activity of commonly used reverse transcriptases[17] to convert these barcode-containing CITE-seq oligonucleotides into cDNA during reverse transcription. We adopted a commonly used streptavidin-biotin interaction to link oligos to antibodies[18] (see methods), and included a disulfide link at the 5' end of the oligonucleotide which allows the oligo to be released from the antibody with reducing agents (Supplementary Fig. 1a). The antibody-oligo complexes are incubated with single-cell suspensions using conditions comparable to staining protocols used in flow cytometry[19]. The cells are washed to remove unbound antibodies and are processed for

scRNA-seq. In this example, we encapsulated single cells into nanoliter-sized aqueous droplets in a microfluidic apparatus designed to perform Drop-seq[1] (Fig. 1b). After cell lysis in droplets, both cellular mRNAs and antibody-derived oligos (released in the reducing conditions of the cell lysis buffer – see Supplementary Fig. 1a) anneal to polyT-containing Drop-seq microparticles (Supplementary Fig. 1b) via their 3' polyA tails. A unique barcode sequence on the oligos attached to the Drop-seq microparticle indexes the cDNA of mRNAs and antibody-oligos of each co-encapsulated cell in the reverse transcription reaction. The amplified antibody-derived tags (hereafter referred to as ADTs) and cDNA molecules can be separated by size and are converted into Illumina-sequencing-ready libraries independently (Supplementary Fig. 1c). Importantly, the two library types are designed to be sequenced together, but because they are generated separately, their relative proportions can be adjusted in a pooled single lane to ensure that the appropriate sequencing depth is obtained for each library.

To assess the ability of our method to distinguish single cells based on surface protein expression, we designed a proof-of-principle 'species-mixing' experiment using. A mixed suspension of human (HeLa) and mouse (4T1) cells was incubated with a mixture of DNA-barcoded anti-mouse and anti-human CD29 (Integrin beta-1) antibodies. After washing to remove unbound antibodies, we performed droplet-based single-cell analysis (Drop-seq[1]) to investigate the concordance between species of origin of the transcripts and ADTs (Fig. 1c-e, Supplementary Fig. 2a). 97.3% of droplets that were identified as containing either human cells (green) or mouse cells (red) by transcriptome alignment (Fig. 1c), were also identified as having the same surface-bound species epitopes using our oligo-labelled antibodies approach (Fig. 1d). Similarly, 99.4% of droplets that were first identified by the antibody-based approach to contain either human or mouse cells showed the same result by transcriptomics analysis. In parallel, we tested the performance of the method in a commercially available system from 10X Genomics and obtained comparable results (Fig. 1f-h, Supplementary Fig. 2b). In both experiments, we deliberately used a high cell concentration to obtain high rates of doublets (droplets containing two or more cells), to demonstrate that we can detect more than one ADT type per droplet, and to correlate mixed-species transcriptome data with mixed-species ADT signals from individual droplets. 96% (Drop-seq) and 89% (10X Genomics) of droplets that

contained human and mouse cell mixtures (Fig. 1c and f, blue) also had ADTs from both human and mouse antibodies (Fig. 1d and g, blue). Cell counts based on RNA or ADT are comparable between both methods (Fig. 1e and h). The consistency between the species classification of cell barcodes based on detected transcripts and ADTs demonstrates the specificity of this method.

We next aimed to benchmark the performance of CITE-seq on mixed complex cell populations. The immune system has been extensively characterized both by cell-surface marker expression[20], and by single-cell gene-expression profiles[6,21,22]. We appended barcodeded oligos to five well-characterized antibodies that recognize specific cell-surface proteins and are routinely used as markers for immune-cell classification: We chose CD3 as a pan-T cell marker, and within the T cell population, CD4 and CD8 to label largely non-overlapping activated T cell subpopulations[20]. CD19 is specifically expressed on B cells and CD56 (together with CD8) marks natural killer (NK) cells[23]. We incubated peripheral blood mononuclear cells (PBMCs) isolated from cord blood with these labeled antibodies and ran the Drop-seq workflow. We measured both cDNA and ADT profiles from 2,616 cells (see methods). Unsupervised graph-based clustering based on RNA expression resulted in partitioning of the identified cell barcodes into recognizable cell types indicated by expression of select marker genes (Fig. 2a, Supplementary Fig. 3), corresponding broadly to T cells, B cells, NK cells, erythroblasts, precursors, and myeloid cells (Fig. 2a, Supplementary Fig. 3).

We first asked whether our detected ADT levels were consistent with our unbiased transcriptomic clustering. For surface markers that are known to be exclusive to individual immune populations, we detected ADT enrichment in the correct populations: CD19 ADTs in the B cell cluster (Fig. 2b), CD56 ADTs in the NK cluster (Fig. 2c) and CD3 ADTs in the T cell cluster (Fig. 2d). We observed that the ADT levels per-cell exhibited stronger enrichment than mRNA-levels of the same genes as ADTs were significantly less prone to ‘drop-out’ events, likely due to the presence of greater copy number of surface proteins. Within the NK cell cluster, we observed an heterogeneity in levels of CD8 ADTs and CD56 ADTs, perhaps reflective of a subdivision of NK-cells (Fig. 2c,f). Indeed, we could construct cytometry-like ‘bi-axial’ gating plots using ADT levels (methods, Supplementary Fig. 4), revealing clear

relationships that are well-established by FACS, including strong anti-correlation of CD4 and CD8 ADT levels, indicative of mutually exclusive expression (Fig. 2g). In fact, we could perform unbiased clustering exclusively based on ADT levels, identifying clear and consistent cell type separation (Supplementary Fig. 5). We conclude that CITE-seq ADT levels represent informative immunophenotypes with increased robustness compared to individual mRNA marker genes.

We next asked whether CITE-seq could enhance our characterization of immune cell phenotypes, compared to scRNA-seq data alone. We noted that our two transcriptomically-derived T cell clusters were separately enriched for CD4 and CD8 ADTs (Fig. 2a, e, f), but did not reflect a clear separation between these two known subsets, likely because CD4 and CD8 are lowly expressed at the RNA level. We therefore re-analyzed our T cell clusters, using only mutually exclusive CD4 and CD8 ADTs, enabling clear separation, and the subsequent identification of transcriptomic markers of CD4 and CD8 cells through differential expression (Fig. 2g, h). This demonstrates that a combined analysis of transcriptome and immunophenotype can enhance our description of cellular phenotypes, particularly when identifying subsets with subtle transcriptomic differences.

Flow cytometry allows for the routine measurement of up to 15 parameters per cell[24,25], largely limited by the separation of emission spectra of fluorophores. In contrast, our molecular barcoding of antibodies is not limited by detection; a modest 10 nucleotide barcode can encode thousands of distinct antibody barcode sequences. This potential multiplexing ability also far exceeds mass cytometry-based parallelization (CyTOF up to 100 tags[25]). Although we have thus far interrogated cell-surface proteins, we envision that mild cell permeabilization and fixation procedures that are used for intracellular antibody staining in signaling-specific FACS assays[25] could also be compatible with RNA sequencing. This may significantly expand the number of markers and biological questions that could be interrogated by this approach. A modified version of CITE-seq in which only ADTs are analyzed on a massively parallel scale without any attempt to capture cellular mRNAs (essentially cytometry by sequencing) can also be envisaged (Supplementary Fig. 6).

CITE-seq allows simultaneous measurement of gene-expression levels and cell-surface epitopes in single cells. This will not only provide a more detailed characterization of cell populations, but also allow the study of post-transcriptional (and post-translational) gene regulation in single cells at an unprecedented depth. The method is highly scalable and limited only by the number of specific antibodies that are available. Importantly, we have shown that the method works on a commercially-available single-cell instrument (10X Genomics) and should be readily adaptable to other droplet, microwell and combinatorial indexing-based high throughput single cell sequencing technologies[2,8-11] with no, or minor customizations. We believe that CITE-seq has the potential to advance single-cell biology by layering an extra dimension on top of single-cell transcriptome data.

## Methods

### Conjugation of Antibodies to DNA-barcoding oligonucleotides

Highly specific, FACS optimized monoclonal antibodies (see below) were conjugated to oligonucleotides containing unique antibody identifier sequences and a polyA tail. Oligonucleotides were conjugated to antibodies by streptavidin-biotin conjugation using the LYNX Rapid Streptavidin Antibody Conjugation Kit (Bio-Rad, USA), according to manufacturer's instructions with modifications. Specifically, we labelled 15µg of antibody with 10µg of streptavidin. At this ratio, up to two streptavidin tetramers can theoretically be conjugated to one antibody, which results in 4-8 binding sites for biotin on each antibody. DNA-oligonucleotides were purchased (IDT, USA) with a 5' biotin modification or with a 5' amine modification and biotinylated using NHS-chemistry according to manufacturer's instructions (EZ Biotin S-S NHS, Thermo Fisher Scientific, USA). The disulfide bond allows separation of the oligo from the antibody with reducing agents. Separation of the oligo from the antibody may not be needed for all applications. Excess Biotin-NHS was removed by gel filtration (Micro Biospin 6, Bio-Rad) and ethanol precipitation. Streptavidin-labelled antibodies were incubated with biotinylated oligonucleotides in excess (1.5x theoretically available free streptavidin) overnight at 4˚C in PBS containing 0.5M NaCl and 0.02% Tween. Unbound oligo was removed from antibodies using centrifugal filters with a 100KDa MW cutoff (Millipore, USA). Removal of

excess oligo was verified by 4% agarose gel electrophoresis (Supplementary Fig. 1a). Antibody-oligo conjugates were stored at 4˚C supplemented with sodium azide and BSA.

### List of Antibodies

- Anti-Human CD29 (Clone MA1-19105) Thermo Fisher, USA
- Anti-Mouse CD29 (Clone MA5-16707) Thermo Fisher, USA
- Anti-Human CD3a (Clone UCHT1) BioLegend, USA (referred to as CD3 throughout the text)
- Anti-Human CD4 (Clone RPA-T4) BioLegend, USA
- Anti-Human CD8a (Clone RPA-T8) BioLegend, USA (referred to as CD8 throughout the text
- Anti-Human CD56 (Clone MEM-188) BioLegend, USA
- Anti-Human CD19 (Clone HIB19) BioLegend, USA

### Antibody-oligo sequences

Depending on the application the PCR-amplification handle in the antibody-barcoding oligos must be changed depending on which sequence read is used for RNA readout (e.g. Drop-seq uses read 2, 10X v1 uses read 1). Our proof-of-principle human/mouse antibody-barcoding oligonucleotide designs included UMIs, which are redundant for Drop-seq and 10X protocols due to the UMI addition to the cDNA at reverse transcription. UMIs may be useful for other iterations of the method where UMIs are not part of the scRNA-seq library preparation protocol.

Species mixing – Drop-seq (containing Nextera read2 handle)

- **BC6: /5AmMC12/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAATNNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- **BC12: /5AmMC12/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGTANNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**

Species mixing – 10X (Single cell 3' version 1, Nextera read1 handle)

- **BC6: /5AmMC12/TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCAATNNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- **BC12: /5AmMC12/TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGTANNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**

PBMC profiling – (v2 Drop-seq oligos, containing TruSeq small RNA read 2 handle)

- v2_BC6: /5AmMC12/**CCTTGGCACCCGAGAATTCCAGCCAATBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- v2_BC12: /5AmMC12/**CCTTGGCACCCGAGAATTCCACTTGTABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- v2_BC1: /5AmMC12/**CCTTGGCACCCGAGAATTCCAATCACGBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- v2_BC5: /5AmMC12/**CCTTGGCACCCGAGAATTCCAACAGTGBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- v2_BC7: /5AmMC12/**CCTTGGCACCCGAGAATTCCACAGATCBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**

- v2_BC2: /5AmMC12/**CCTTGGCACCCGAGAATTCCACGATGTBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
- v2_BC4: /5AmMC12/**CCTTGGCACCCGAGAATTCCATGACCABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**

**Cell DNA-barcoded antibody labelling for CITE-seq**

Cells were resuspended in cold PBS containing 2%BSA and 0.01%Tween (PBT) and filtered through 40µm cell strainers (Falcon, USA) to remove potential clumps and large particles. Cells were then incubated for 10 minutes with Fc receptor block (TruStain FcX, BioLegend, USA) to block non-specific antibody binding. Subsequently cells were incubated in with mixtures of barcoded antibodies for 30 minutes at 4˚C. Cells were washed 3x by resuspension in PBS containing 2%BSA and 0.01%Tween, followed by centrifugation (~480x*g* 5 minutes) and supernatant exchange. After the final wash cells are resuspended at appropriate cell concentration for Drop-seq[1] or 10X Genomics[3] applications.

**CITE-seq procedure**

In addition to the canonical RNA-dependent DNA polymerase activity, commonly used reverse transcriptases have also DNA-dependent DNA polymerase activity. This activity allows our DNA-barcoding oligonucleotides to be copied into cDNA during reverse transcription, generating ADTs. Specific protocol details for CITE-seq applied to Drop-seq and 10X are outlined below.

**Drop-seq – CITE-seq**

Drop-seq was performed as described[1] with modifications. For the human/mouse mixing experiment cells were loaded at a concentration of 400 cells/µL to achieve a high doublet rate. For PMBC experiments cells were loaded at 150 cells/µL. cDNA was amplified for 10 cycles and products were then size separated with Ampure Beads (Beckman Coulter, USA) into <300 nt fragments containing antibody derived tags (ADTs) and >300 nt fragments containing cDNAs derived from cellular mRNA. ADTs were amplified 10 additional cycles using specific primers that append P5 and P7 sequences for clustering on Illumina flowcells. Alternatively, antibody tags can be amplified directly from thoroughly washed Drop-seq beads after RNA-cDNA amplification using specific primers for the antibody oligo and Drop-seq bead-RT oligo. cDNAs derived from RNA were converted into sequencing libraries by tagmentation as described[1]. Both libraries were quantified and merged at appropriate concentrations and

sequenced on a HiSeq 2500 Rapid Run with v2 chemistry per manufacturer's instructions (Illumina, USA).

**10X – CITE-seq**

The 10X single cell run was performed according to the manufacturer's instructions (Single Cell 3' v1, 10X Genomics, USA) with modifications. 10,000 cells were loaded to yield an intermediate/high doublet rate. cDNA was amplified for 10 cycles and products were then size separated with Ampure Beads (Beckman Coulter, USA) into <300 nt fragments containing antibody derived tags (ADTs) and >300 nt fragments containing cDNAs derived from cellular mRNA. ADTs were amplified 10 additional cycles using specific primers that append P5 and P7 sequences for clustering on Illumina flowcells. A sequencing library from cDNAs derived from RNA was generated using a tagmentation based approach akin to that used in Drop-seq. Both libraries were merged and sequenced on a HiSeq Rapid Run according to manufacturer's instructions (Illumina, USA).

**Cell culture**

HeLa (human) and 4T1 (mouse) cells were obtained from ATCC and maintained according to standard procedures in Dulbecco's Modified Eagle's Medium (Thermo Fisher, USA) supplemented with 10% fetal bovine serum (Thermo Fisher, USA) at 37˚C with 5% $CO_2$. For the species mixing experiment, cells were mixed in equal proportions and incubated with DNA barcoded CITE-seq antibodies as described above.

**PBMC isolation**

Peripheral blood mononuclear cells were isolated from cord blood (New York Blood Center) as described[26]. Cells were kept on ice during and after isolation.

## Computational methods

**Single cell RNA data processing and filtering**

The raw Drop-seq data was processed with the standard pipeline (Drop-seq tools version 1.12 from McCarroll lab). 10X data was processed using Cell Ranger 1.2 using default parameters.

Reads were aligned to the human reference sequence GRCh37/hg19 (PBMCs), or an hg19 and mouse reference mm10 concatenation for species mixing experiments.

No further filtering was applied to 10X data. Drop-seq data of the species mixing experiment was filtered to contain only cells with at least 500 UMIs mapping to human genes, or 500 UMIs mapping to mouse genes. For the PBMCs, we kept only cells with PCT_USABLE_BASES >= 0.5 (fraction of bases mapping to mRNA, this is part of the metrics outputted by the default processing pipeline). We further removed any cells with less than 300 genes detected and cells with a total number of UMIs that is more than 5 standard deviations above or below the mean.

**Single cell ADT data processing and filtering**

Antibody and cell barcodes were directly extracted from the reads in the fastq files. Since the antibody barcodes were sufficiently different in the species mixing experiment, we also counted sequences with Hamming distance less than 4. For the PBMCs we counted sequences with Hamming distance less than 2. Reads with the same combination of cellular, molecular and antibody barcode were only counted once.

We kept only cells that had at least 10 ADT counts and also passed the RNA-specific filters.

**PBMC RNA normalization and clustering**

We converted the matrix of UMI counts into a log-normalized expression matrix $x$ with $x_{i,j} = \log\left(\frac{c_{i,j}*10{,}000}{m_j}\right)$, where $c_{i,j}$ is the molecule count of gene $i$ in cell $j$ and $m_j$ is the sum of all molecule counts for cell $j$. After normalization each gene was scaled to have mean expression 0 and variance 1.

We identified 1448 highly variable genes by fitting a smooth line (LOESS, span=0.33, degree=2) to log10(var(UMIs)/mean(UMIs)) as a function of log10(mean(UMIs)) and keeping all genes with a standardized residual above 0.25 and a detection rate of at least 1%.

To cluster the cells, we performed dimensionality reduction followed by modularity optimization. We ran principal component analysis (PCA) using the expression matrix of variable genes. To determine the number of significant dimensions, we looked at the percent change in successive eigenvalues. The last eigenvalue to feature a reduction of at least 10% constituted our significant number of dimensions (in this case the number was 6). For

clustering we used a modularity optimization algorithm that finds community structure in the data[27] with Jaccard similarities (neighborhood size 52 which is 2% of the cells, Euclidean distance of significant PCs) as edge weights between cells. For 2D visualization we further reduced the dimensionality of the data to 2 using t-SNE[28,29].

**BMC ADT normalization and clustering**

Since each ADT count for a given cell can be interpreted as part of a whole (all ADT counts assigned to that cell), and there are only 5 components in this experiment, we treated this data type as compositional data and applied the centered logratio (CLR) transformation(Aitchison 1989). Explicitly, we generated a new CLR-transformed ADT vector $y$ for each cell where $y = clr(x) = [\ln\left(\frac{x_1}{g(x)}\right), \ln\left(\frac{x_2}{g(x)}\right), \ldots, \ln\left(\frac{x_5}{g(x)}\right)]$ and $x$ is the vector of ADT counts including one pseudocount for each component, and $g(x)$ is the geometric mean of $x$.

We noticed that the ADT counts were on slightly different scales for the different antibodies, perhaps due to differences in antibody specificity and/or epitope abundance. To compensate for the resulting shifts in the non-specific baseline ADT signal, we examined the density distribution of the CLR-transformed ADT counts of all antibodies (Supplementary Fig. 4a). We fitted each univariate density as a Gaussian mixture of 3 components with equal variance. The component with largest mean defines the 'on' state where protein is present, and the CLR-transformed ADT cutoff is the smallest value that is assigned to the 'on' state.

To cluster cells based on ADT counts, the same general approach as for the RNA data was taken, except no dimensionality reduction was performed. Instead the CLR-transformed ADT counts were scaled to mean 0 and variance 1 for each antibody. Cell-to-cell weights were Jaccard similarities based on Euclidean distance and neighborhood size 150 (5% of the cells).

**PBMC differential expression test**

We wanted to identify genes that were differentially expressed between the two T-cell clusters, as well as between the CD4+ and CD8+ T-cells. In the first case we directly compared all 467 cells in cluster 7 to all 621 cells in cluster 8. In the second case we subsetted all cells that were in either cluster 7 or 8 into two groups based on CLR-transformed ADT counts and the antibody specific cutoffs defined above: 1) 694 cells that were CD4+ and CD8- and 2) 328 cells that were CD4- and CD8+. There were 56 cells in cluster 7 or 8 that were CD4- and CD8-, and

10 cells that were both CD4+ and CD8+. These cells were not used in the second differential expression test. The test method was based on the following negative binomial regression model. For a gene $i$ and its vector of UMI counts $c_i$ and a group indicator variable $g$ we fit the two models:

Model 1: $\log\big(E(c_i)\big) = \alpha + \beta_1 m + \epsilon$

Model 2: $\log\big(E(c_i)\big) = \alpha + \beta_1 m + \beta_2 g + \epsilon,$

with the sequencing-depth related factor $m$, the number of total UMI counts per cell. The comparison of the two models using a likelihood ratio test determines the p-value of model 2 providing a better fit. The log-fold change is directly given by the coefficient of the group indicator variable as $\log_2(e^{\beta_2})$.

**Acknowledgements**

We acknowledge members of the innovation lab S. Jaini and K. Pandit for critical discussions and support. We thank M. Coppo and S. Pescatore at NYGC for sequencing support. CH was supported by a Deutsche Forschungsgemeinschaft research fellowship.

**Competing financial interests**

The authors declare no competing financial interests.

## REFERENCES


1. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *CELL* **161,** 1202–1214 (2015).
2. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *CELL* **161,** 1187–1201 (2015).
3. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8,** 1–12 (2017).
4. Pontén, F. *et al.* A global view of protein expression in human cells, tissues, and organs. *Mol Syst Biol* **5,** 1–9 (2009).
5. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *CELL* **163,** 1663–1677 (2015).
6. Wilson, N. K. *et al.* Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *STEM* **16,** 712–724 (2015).
7. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128,** e20–31 (2016).
8. Yuan, J. & Sims, P. A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Nature Publishing Group* **6,** 33883 (2016).
9. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Meth* (2017). doi:10.1038/nmeth.4179
10. Cao, J. *et al.* Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. 1–35 (2017). doi:10.1101/104844
11. Rosenberg, A. B. *et al.* Scaling single cell transcriptomics through split pool barcoding. 1–13 (2017). doi:10.1101/105163
12. Ståhlberg, A., Thomsen, C., Ruff, D. & Åman, P. Quantitative PCR analysis of DNA, RNAs, and proteins in the same single cell. *Clinical Chemistry* **58,** 1682–1691 (2012).
13. Genshaft, A. S. *et al.* Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* **17,** 188 (2016).
14. Albayrak, C. *et al.* Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Molecular Cell* **61,** 914–924 (2016).
15. Darmanis, S. *et al.* Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *CellReports* **14,** 380–389 (2016).

16. Frei, A. P. *et al.* Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat Meth* **13,** 269–275 (2016).
17. Baranauskas, A. *et al.* Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. *Protein Eng. Des. Sel.* **25,** 657–668 (2012).
18. Adler, M., Wacker, R. & Niemeyer, C. M. Sensitivity by combination: immuno-PCR and related technologies. *Analyst* **133,** 702–18 (2008).
19. Wood, B. Multicolor immunophenotyping: human immune system hematopoiesis. *Methods Cell Biol.* **75,** 559–576 (2004).
20. Murphy, K., Travers, P. & Walport, M. *Janeways Immuno Biology 7th ed. Garland Pub*. (Inc. New York and London, 2008).
21. Zheng, G. X. Y. *et al. Massively parallel digital transcriptional profiling of single cells*. *bioRxiv* 1–46 (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/065912
22. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347,** 1258367–1258367 (2015).
23. Campbell, J. P., Guy, K., Cosgrove, C., Florida-James, G. D. & Simpson, R. J. Total lymphocyte CD8 expression is not a reliable marker of cytotoxic T-cell populations in human peripheral blood following an acute bout of high-intensity exercise. *Brain, Behavior, and Immunity* **22,** 375–380 (2008).
24. Chattopadhyay, P. K. & Roederer, M. Cytometry: today‘s technology and tomorrow’s horizons. *Methods* **57,** 251–258 (2012).
25. Bendall, S. C. & Nolan, G. P. From single cells to deep phenotypes in cancer. *Nat Biotechnol* 1–9 (2012). doi:10.1038/nbt.2283
26. Breton, G. E. L., Lee, J., Liu, K. & Nussenzweig, M. C. Defining human dendritic cell progenitors by multiparametric flow cytometry. *Nature Protocols* **10,** 1407–1422 (2015).
27. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008,** P10008 (2008).
28. Maaten, L. V. D. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).
29. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* (2014).

# Figure 1

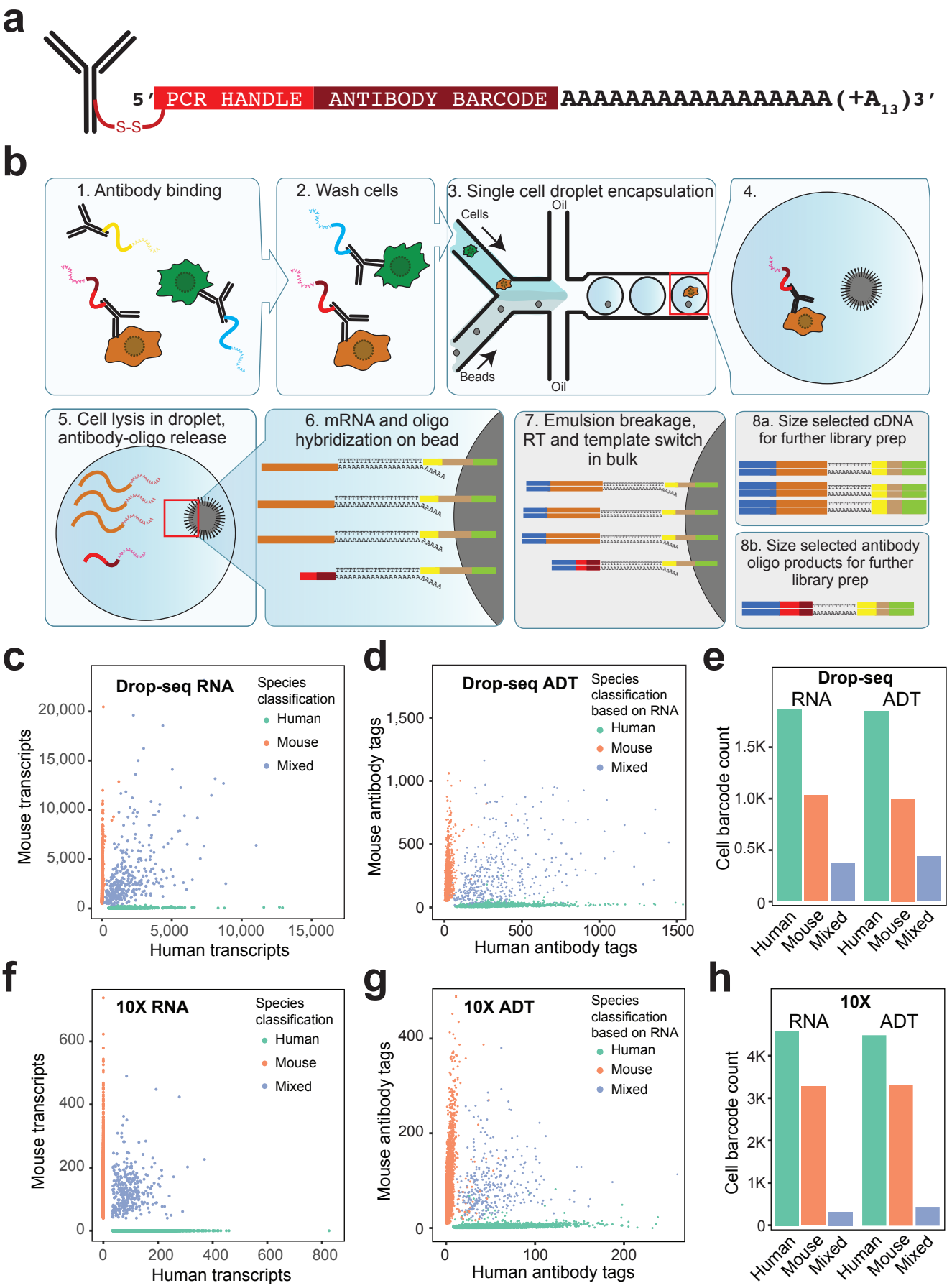

a
5′ PCR HANDLE ANTIBODY BARCODE AAAAAAAAAAAAAAAAAA(+A13)3′
S-S
b
1. Antibody binding
2. Wash cells
3. Single cell droplet encapsulation
Cells
Oil
Beads
Oil
4.
5. Cell lysis in droplet, antibody-oligo release
6. mRNA and oligo hybridization on bead
7. Emulsion breakage, RT and template switch in bulk
8a. Size selected cDNA for further library prep
8b. Size selected antibody oligo products for further library prep
c
Drop-seq RNA
Species classification
Human
Mouse
Mixed
Mouse transcripts
Human transcripts
d
Drop-seq ADT
Species classification based on RNA
Human
Mouse
Mixed
Mouse antibody tags
Human antibody tags
e
Drop-seq
RNA
ADT
Cell barcode count
Human
Mouse
Mixed
f
10X RNA
Species classification
Human
Mouse
Mixed
Mouse transcripts
Human transcripts
g
10X ADT
Species classification based on RNA
Human
Mouse
Mixed
Mouse antibody tags
Human antibody tags
h
10X
RNA
ADT
Cell barcode count
Human
Mouse
Mixed

**Figure 1 | CITE-seq enables simultaneous detection of single cell transcriptomes and protein markers.** (**a**) Schematic overview of the DNA-barcoded antibodies used in CITE-seq. Antibodies are linked to oligonucleotides containing an amplification handle, a unique antibody identifier barcode and a polyA tail. A disulfide bond between the antibody and oligo allows release. (**b**) Schematic representation of CITE-seq in combination with Drop-seq[1]. Cells are incubated with antibodies (1). Cells are washed (2) and passed through a microfluidic chip (3) where a single cell and one bead are occasionally encapsulated in the same droplet (4). After cell lysis (5) mRNAs and released antibody-oligos anneal to oligos on Drop-seq beads (6). Reverse transcription and template switch is performed in bulk after droplet emulsion breakage (7). After amplification, full length cDNA and antibody-oligo products can be separated by size and amplified independently (8) (**c - h**) Analysis of mixtures of mouse and human cells that were incubated with oligo-tagged-antibodies specific for either human or mouse cell-surface markers (integrin beta CD29) and passed through Drop-seq (c-e) or 10X Genomics (f-h) workflows. In both experiments, we deliberately used a high cell concentration to obtain high rates of doublets (droplets containing two or more cells), to correlate mixed-species transcriptome data to mixed-species protein signals from individual droplets. (**c** and **d**) quantification of the number of human and mouse transcripts associating to each cell barcode. Each green point indicates a cell barcode (i.e. droplet containing one or more cells) from which we measured >90% human transcripts; each red point indicates a cell barcode with >90% mouse transcripts. Blue points indicate cell barcodes (i.e. droplets) from which we observed a mixture of human and mouse transcripts. (**d** and **g**) Quantitation of species-specific antibody tags (ADTs) associated with each cell barcode. Points are colored based on species classifications using transcripts in (c) and (d), respectively. (**e** and **h**) Quantification of human, mouse or mixed-cell barcodes based on RNA transcripts, or ADTs, in Drop-seq and 10X Genomics.

# Figure 2

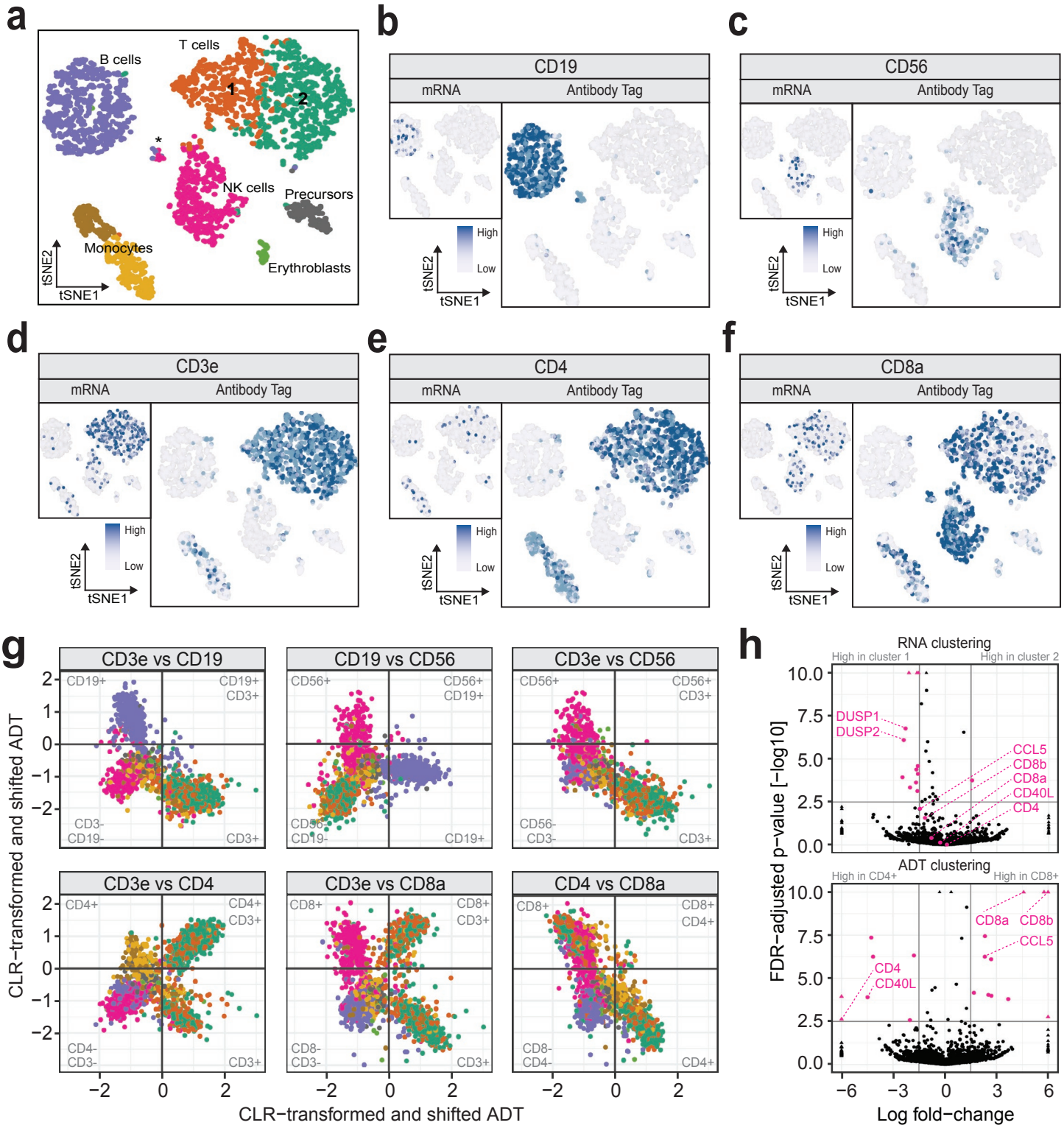

a
B cells
T cells
1
2
*
NK cells
Precursors
Monocytes
Erythroblasts
tSNE2
tSNE1
b
CD19
mRNA
Antibody Tag
High
Low
c
CD56
d
CD3e
e
CD4
f
CD8a
g
CD3e vs CD19
CD19+
CD19+ CD3+
CD3- CD19-
CD3+
CD19 vs CD56
CD56+
CD56+ CD19+
CD56- CD19-
CD19+
CD3e vs CD56
CD56+ CD3+
CD56- CD3-
CD3e vs CD4
CD4+
CD4+ CD3+
CD4- CD3-
CD3e vs CD8a
CD8+
CD8+ CD3+
CD8- CD3-
CD4 vs CD8a
CD8+ CD4+
CD8- CD4-
CLR−transformed and shifted ADT
h
RNA clustering
High in cluster 1
High in cluster 2
DUSP1
DUSP2
CCL5
CD8b
CD8a
CD40L
CD4
ADT clustering
High in CD4+
High in CD8+
CD8a
CD8b
CCL5
CD4
CD40L
FDR−adjusted p−value [−log10]
Log fold−change

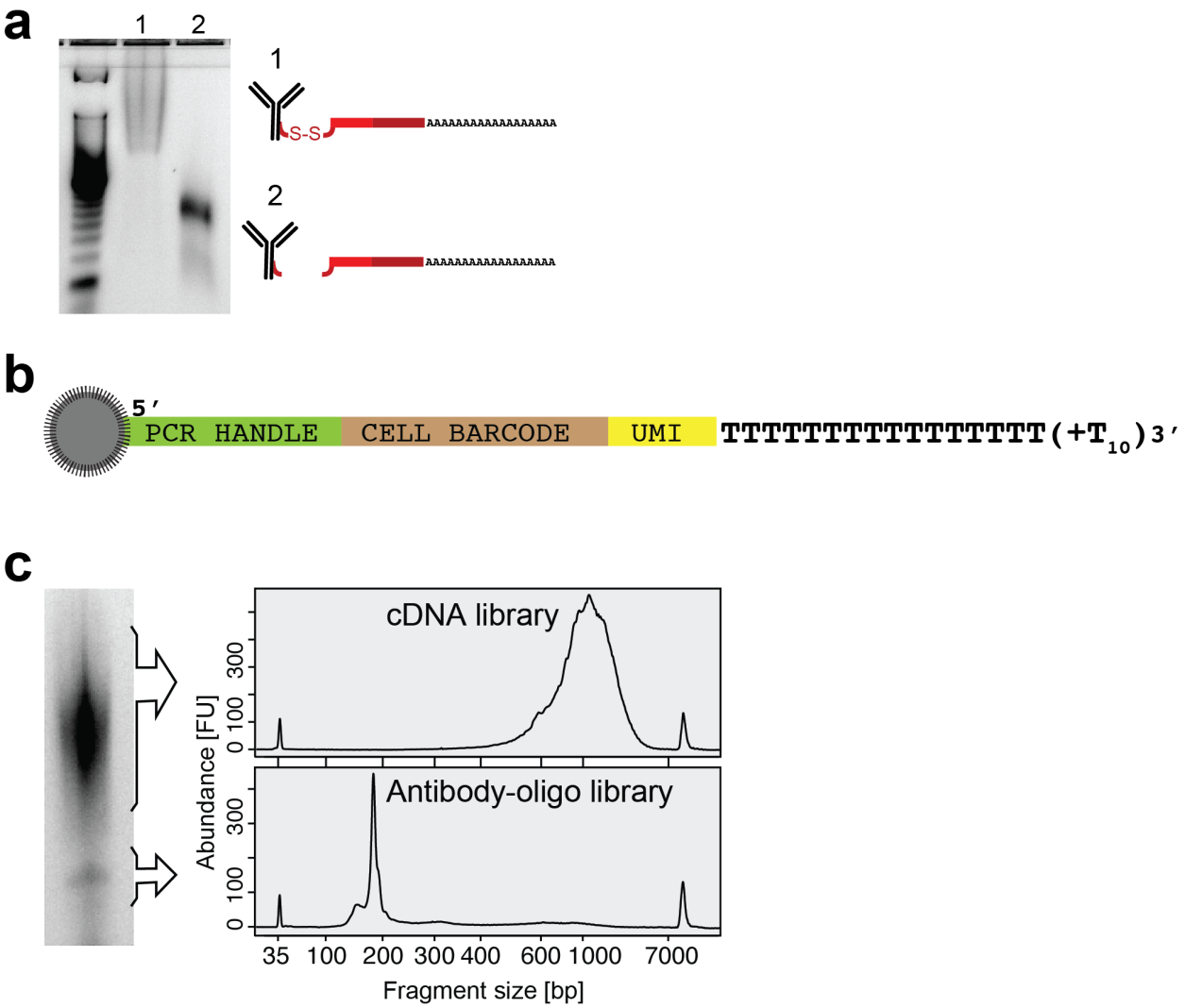
**Figure 2 | CITE-seq allows for detailed classification of peripheral blood mononuclear cells.** (**a**) Clustering of 2,616 CITE-seq single-cell expression profiles of PBMCs into 8 cell populations. The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among all cells (methods). Major cell types in peripheral blood can be discerned based on marker gene expression (Supplementary Fig. 3). T cells (2 clusters), B cells, NK cells, monocytes (2 clusters), precursors, erythrocytes, and putative doublets co-expressing multiple lineage markers (*) are indicated. (**b-f**) mRNA and corresponding ADT signal for CD3e, CD4, CD8a, CD56 and CD19 projected on the tSNE plot from (a). (**g**) Pairwise comparison of different ADT levels in single cells. Colors correspond to cell clusters in panel (a). ADT counts were centered logratio (CLR) transformed and antibody-specific cutoffs were determined by modeling the density distribution and subtracted (see methods, Supplementary Fig. 4a). Within each panel, the ADT mentioned first corresponds to the X axis. (**h**) Differential gene expression analysis of cells in the two T-cell clusters in panel (a) based on RNA data (top panel) and based on CD4 and CD8 ADT levels (lower panel). CD4 and CD8 T-cell specific marker genes were identified as differentially expressed when dividing the T-cells using ADT data, but not with RNA-seq data. For better visualization, genes with large absolute fold change were clipped to -6 or 6, and small p–values were clipped to $10^{-10}$. These genes are indicated by triangles instead of circles.

# Supplementary Table 1

| Technology | # transcripts | # proteins | # cells | Reference |
|---|---|---|---|---|
| Single-cell RNA-seq (droplet based) | > 4,000 | - | > 10,000 | 1, 2, 3 |
| CyTOF | - | < 100 | > 10,000 | 25 |
| FACS | - | < 20 | > 10,000 | 24 |
| Index-sorting RNA-seq | > 4,000 | < 20 | ~ 96 - hundreds | 5, 6, 7 |
| Wafergen ICELL8 | > 4,000 | ~ 4 | ~ 1,800 | |
| Fluidigm C1 | > 4,000 | ~ 4 | ~ 96 - ~ 800 | |
| PLA/PEA & qPCR | ~ 96 | ~ 36 | ~ 96 - hundreds | 12, 13, 14, 15 |
| PLAYR | ~ 20 | ~ 20 | > 10,000 | 16 |

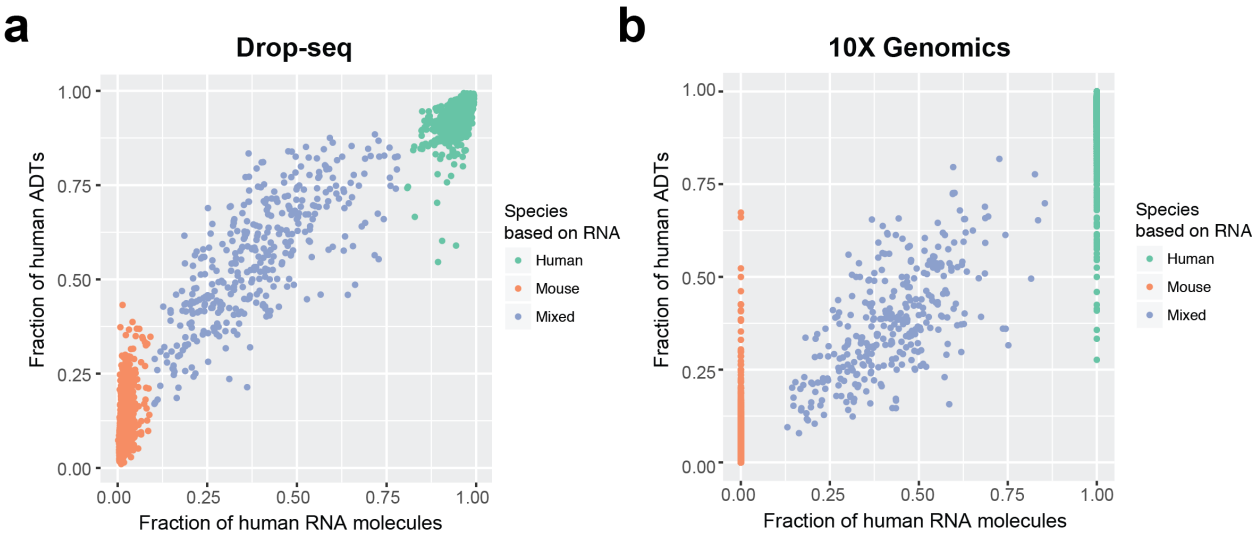
**Supplementary Table 1 | Overview of methods that allow multiplexed measurements of RNA and/or proteins in single cells.**

# Supplementary Figure 1




**Supplementary Figure 1 | CITE-seq library preparation.** (**a**) Antibody-oligonucleotide complexes appear as a high-molecular-weight smear when run on an agarose gel (1). Cleavage of the oligo from the antibody by reduction of the disulfide bridge collapses the smear to oligo length (2). (**b**) Drop-seq beads are microparticles containing oligonucleotides with a common PCR handle, a unique cell barcode, followed by a unique molecular identifier (UMI) and a polyT tail{Macosko:2015ht}. (**c**) Reverse transcription and amplification produces two product populations with distinct sizes (right panel). These can be size separated and amplified independently to obtain full length cDNAs (top panel, capillary electrophoresis trace) and ADTs (bottom panel, capillary electrophoresis trace).

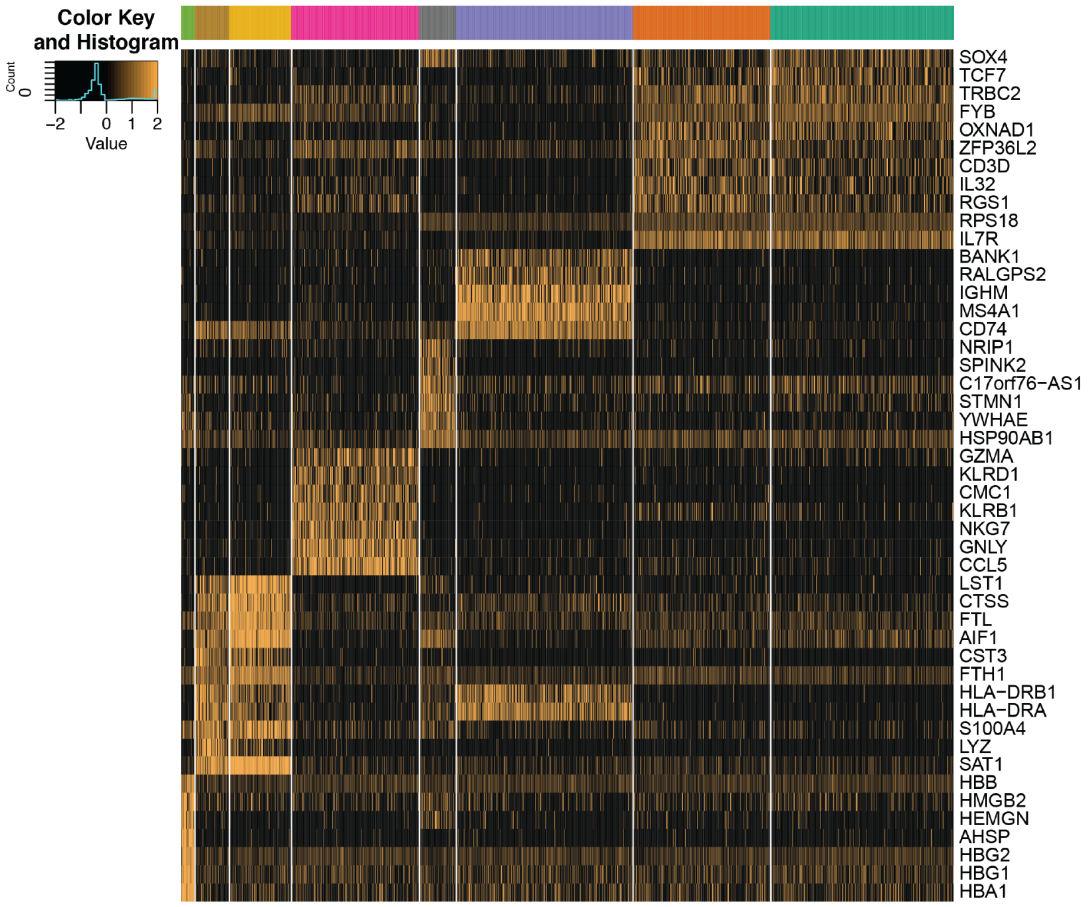
# Supplementary Figure 2



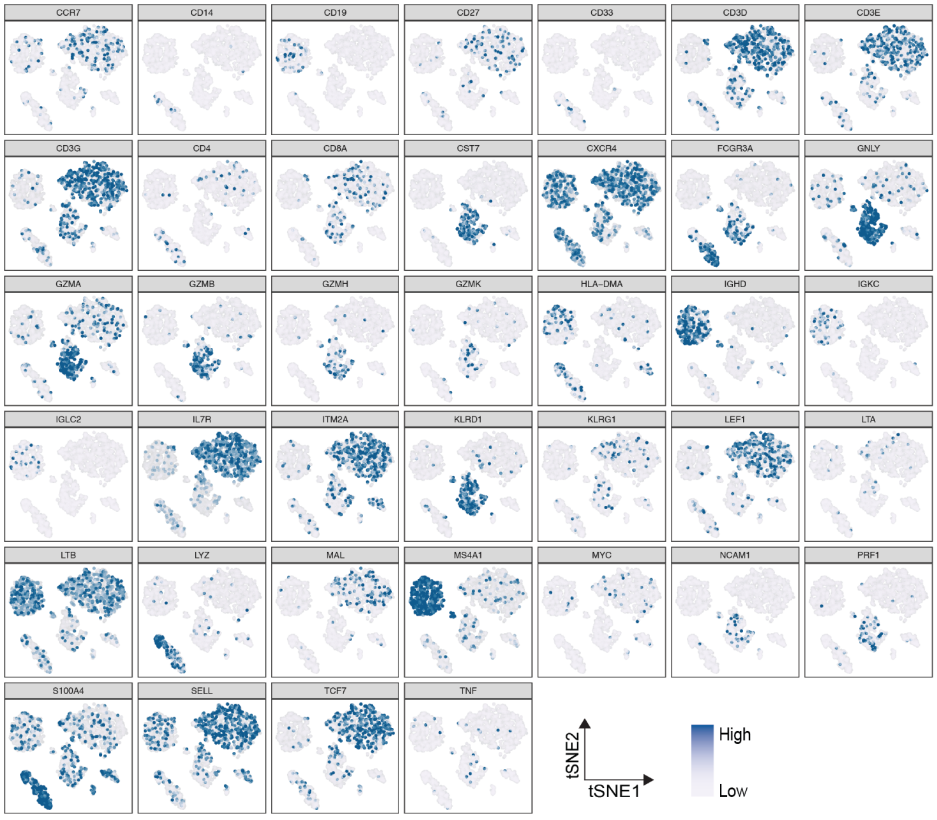

**Supplementary Figure 2 | Human and mouse RNA and ADT fractions. (a-b)** Analysis of mixtures of mouse (4TI) and human (HeLa) cells that were incubated with oligo-tagged-antibodies specific for human or mouse cell-surface markers (integrin beta, CD29) and passed through the Drop-seq (**a**) or 10X Genomics (**b**) workflow. Each point represents one cell barcode (*i.e.,* droplet containing one or more cells). Fractions of human RNA molecules compared to human ADTs for detected cell barcodes.

# Supplementary Figure 3

**a**

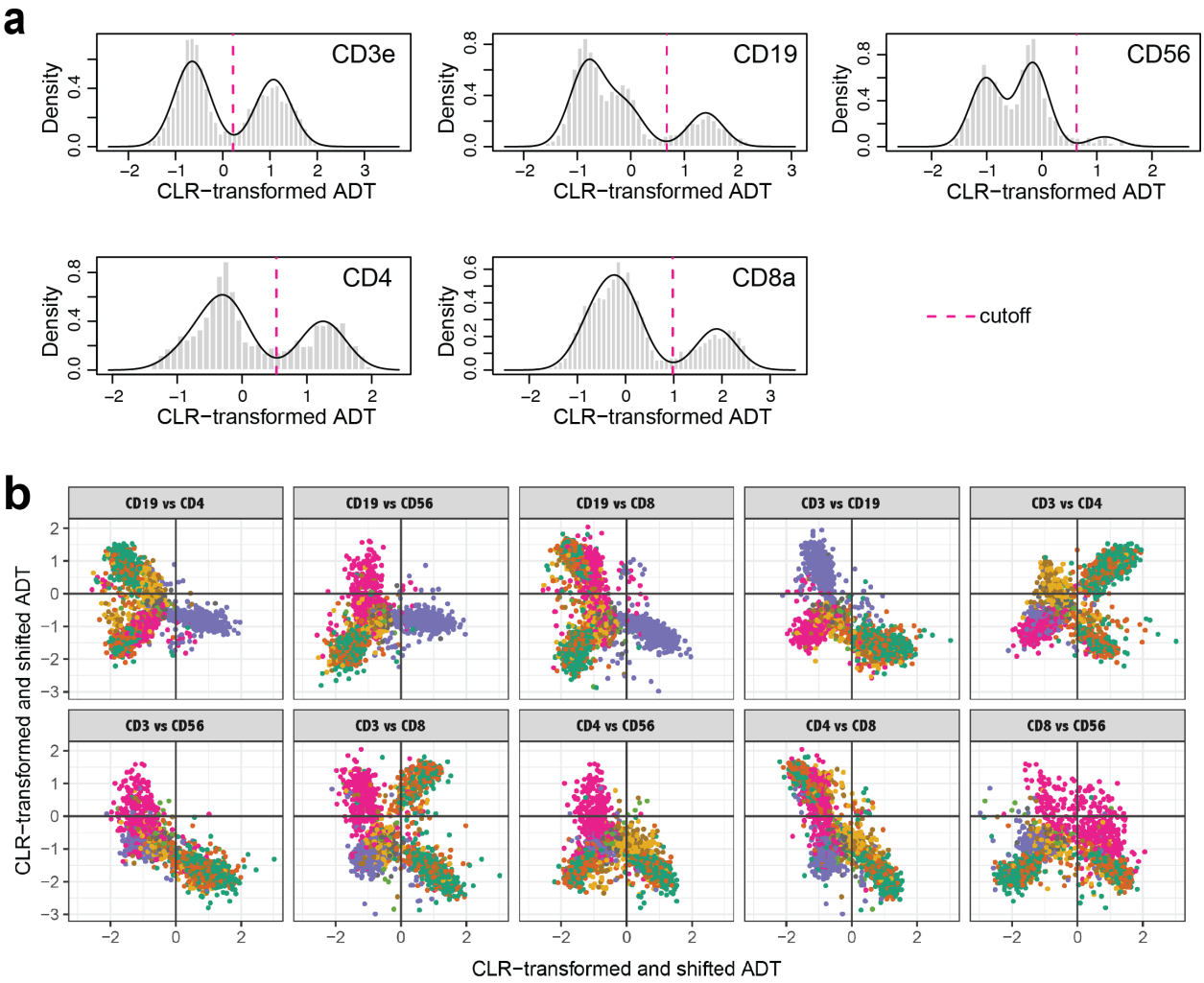



**b**



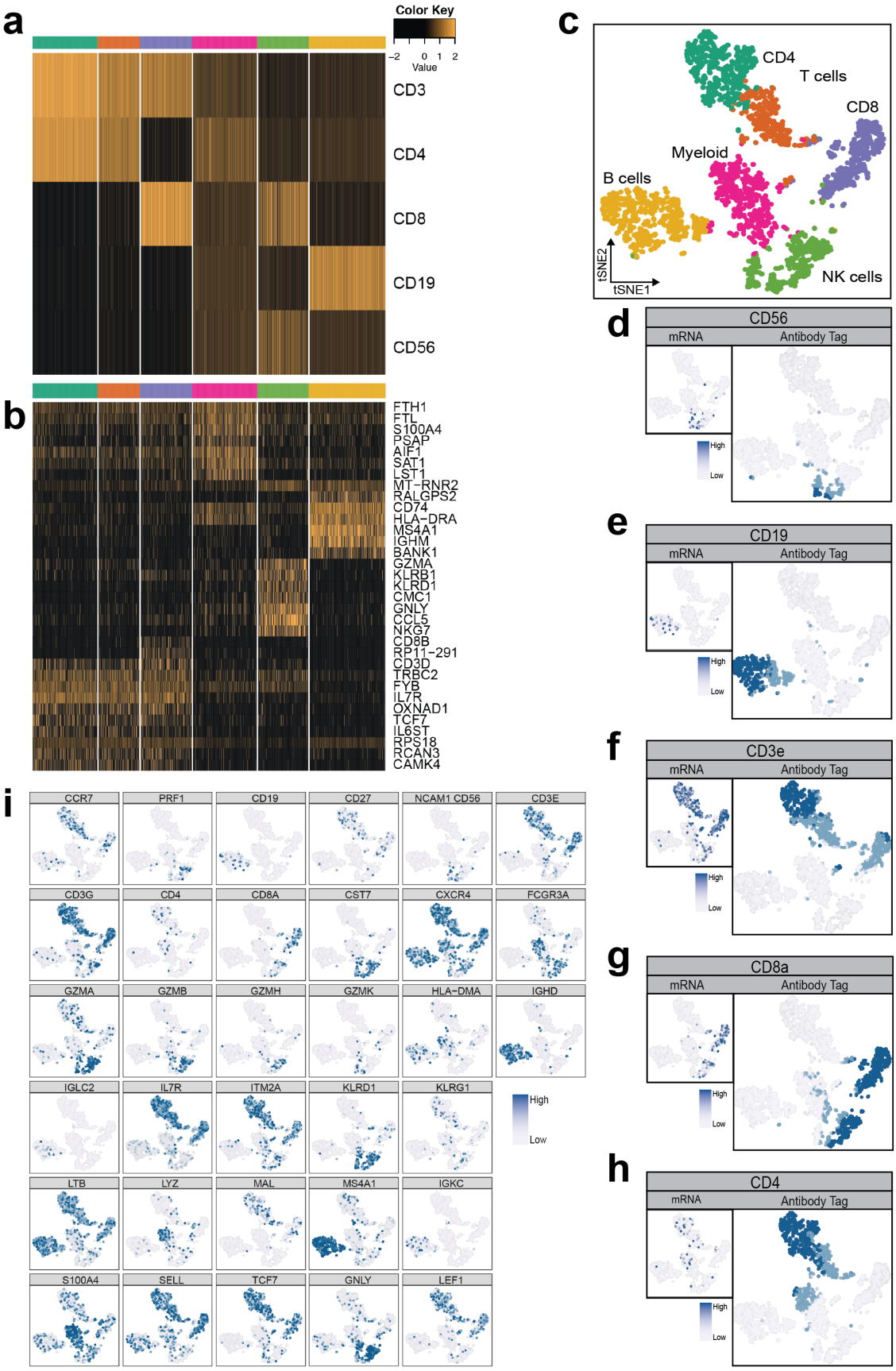
**Supplementary Figure 3 | Marker gene expression in clustered PBMC datasets.** (**a**) Gene expression heatmap of most differentially expressed marker genes used to define clusters. Dimensionality reduction followed by modularity optimization was used to cluster 2,616 PBMCs (methods). Cluster color assignments are identical to Figure 2a. (**b**) Expression of individual marker genes in the context of tSNE representation of cell relationships based on single-cell gene-expression profiles. Levels of transcripts corresponding to specific marker genes are indicated by blue shading. Plots were used to define cell types in different clusters.

# Supplementary Figure 4




**Supplementary Figure 4 | Resolving single-cell RNA clusters using ADTs.** (**a**) Histograms of CLR-transformed ADT counts. Solid line shows density of 3-component Gaussian mixture model with equal variances (methods). The smallest value of a cell assigned to the right-most component is the cutoff for presence of an ADT signal. (**b**) Pairwise comparison of different ADT levels. Colors based on single-cell clustering shown in Figure 2a. Pairs of CLR-transformed ADT values after subtraction of cutoff (methods). Within each panel, the ADT mentioned first corresponds to the X axis.

# Supplementary Figure 5



**Supplementary Figure 5 | Clustering of PBMCs based on ADTs**. (**a**) CITE-seq single-cell ADT profiles of PBMCs were clustered using modularity optimization resulting in 6 cell populations with distinct antibody compositions. (**b**) Single-cell gene-expression heatmap of top marker genes of ADT-derived clusters. Colors of clusters in top panel represent colors in (a). (**c**) Two-dimensional representation (tSNE) of cell relationships based on single-cell ADT profiles. Cluster colors are represented in (a) and (b). (**d – h**) Comparison of mRNA levels and ADT levels of markers in clusters defined by ADT levels. (**i**) Expression of individual marker genes in the context of tSNE representation of cell relationships based on ADT levels. Levels of transcripts corresponding to specific marker genes are indicated by blue shading. Plots were used to define cell types in different clusters.

# Supplementary Figure 6

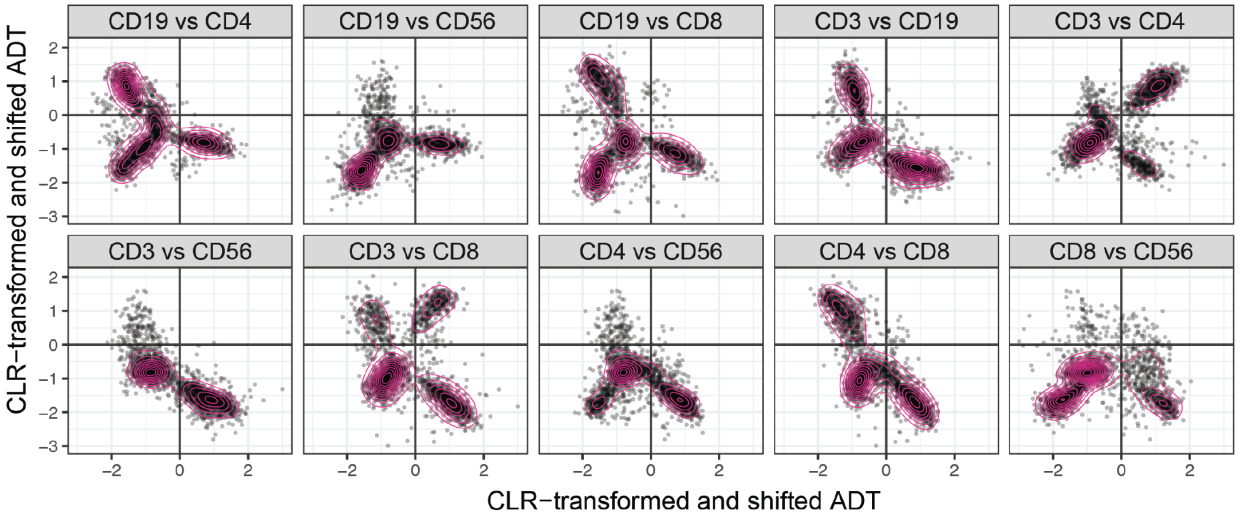



**Figure S6 | Cytometry by sequencing.** Bi-axial gating plots of pairs of CLR-transformed ADTs after subtraction of antibody-specific cutoff. 2,616 cells are plotted as in Supplementary Figure 4b. Within each panel, the ADT mentioned first corresponds to the X axis. Large numbers of cells can be profiled at low sequencing depth when only ADTs without single-cell transcriptomes are profiled. Cells can then be analyzed based only on their immunophenotype to essentially achieve cytometry by sequencing at a highly multiplexed scale.