HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 January 31.

Published in final edited form as:

Nat Methods. 2017 September ; 14(9): 865–868. doi:10.1038/nmeth.4380.

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

Marlon Stoeckius[1,*], **Christoph Hafemeister**[1], **William Stephenson**[1], **Brian Houck-Loomis**[1], **Pratip K. Chattopadhyay**[3], **Harold Swerdlow**[1], **Rahul Satija**[1,2], and **Peter Smibert**[1]

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

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

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

Abstract

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 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, detailed definitions of cellular states were routinely obtained via carefully curated panels of fluorescently labeled antibodies directed at cell surface proteins, which are often reliable indicators of cellular activity and function[4]. Recent studies[5,6] have demonstrated the potential for coupling 'index-sorting' measurements with single-cell

---

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

[*]Correspondence: mstoeckius@nygenome.org.
ORCIDs:
Marlon orcid.org/0000-0002-5658-029X
Christoph orcid.org/0000-0001-6365-8254
William orcid.org/0000-0002-3779-417X
Pratip orcid.org/0000-0002-5457-9666
Peter orcid.org/0000-0003-0772-1647
Rahul orcid.org/0000-0001-9448-8833
Harold orcid.org/0000-0002-9510-160X
Brian orcid.org/0000-0002-1863-1199

Author contributions

MS conceived and designed the study with input from BH-L, RS, HS and PS. MS performed all experiments. CH and RS designed and contributed the computational analyses. WS assisted with Drop-seq experiments. PC provided conceptual input on how to benchmark CITE-seq to flow cytometry and performed multiparameter flow cytometry analysis. MS, CH, RS and PS interpreted the data. MS and PS wrote the manuscript with input from all authors.

Competing financial interests

MS, BH-L and PS have filed a patent application based on this work.

Data availability

All raw data generated in this project has been deposited to the gene expression omnibus (GEO) with the accession code XXXXXX.

transcriptomics, enabling the mapping of immunophenotypes onto transcriptomically-derived clusters. However, massively parallel approaches based on droplet microfluidics[1–3], microwells[7,8] or combinatorial indexing[9,10] do not utilize cytometry for cellular isolation, and therefore cannot couple protein information with cellular transcriptomes, representing a significant limitation for these approaches. Targeted methods to simultaneously measure transcripts and proteins in single cells are limited in scale and/or can only profile a few genes and proteins in parallel[11–15] (Supplementary Table 1).

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 cellular phenotypes than scRNA-seq alone.

We hypothesized that a DNA oligonucleotide conjugated to an antibody could be measured by sequencing as a digital readout of protein abundance. We conjugated antibodies to oligonucleotides designed to 1) be captured by oligo dT-based RNA sequencing library preparations, 2) contain a barcode sequence for identification the antibody and 3) allow subsequent specific amplification by PCR (Fig 1a). We adopted a commonly used streptavidin-biotin interaction to link the 5' end of oligos to antibodies, and included a disulfide link, which allows the oligo to be released from the antibody in reducing conditions (Supplementary Fig. 1a). The antibody-oligo complexes are incubated with single-cell suspensions using conditions comparable to staining protocols used in flow cytometry, after which cells are washed to remove unbound antibodies and 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 anneal to polyT-containing Drop-seq microparticles (Supplementary Fig. 1b,c) 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 converted into Illumina-sequencing-ready libraries independently (Supplementary Fig. 1c,d). 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, leveraging a species-specific and highly expressed protein marker CD29 (Integrin beta-1). A mixed suspension of human (HeLa) and mouse (4T1) cells was incubated with a mixture of DNA-barcoded anti-mouse and anti-human CD29 antibodies. After washing to remove unbound antibodies, we performed Drop-seq[1] to investigate the concordance between species of origin of the transcripts and ADTs (Fig. 1c–e, Supplementary Fig. 2a). We deliberately used a high cell concentration to obtain high rates of multiplets (droplets containing two or more cells), to correlate mixed-species transcriptome data with mixed-species ADT signals from individual droplets. 97.2% of droplets that were identified as having contained either human, mouse, or mixed cells by transcriptome (Fig. 1c), had the same species classification by ADT counts (Fig. 1d). Cell counts based on RNA or ADT are highly correlated between both methods (Fig. 1e), demonstrating low drop-out rate of ADT signals. We performed the same experiment using a commercially available system from 10× Genomics and obtained comparable results (Supplementary Fig. 2b–e).

We next wanted to characterize the quantitative nature of our CITE-seq ADT protein readout. Flow cytometry is the gold-standard for identification and enumeration of cell subsets based on quantitative differences in surface markers[16,17]. We therefore aimed to benchmark the sensitivity of CITE-seq protein detection to flow cytometry using immune cells as a model system. We prepared a set of CITE-seq antibodies directed against markers commonly used in flow cytometry to identify and discriminate relevant immune sub-populations. We performed multiparameter flow cytometry (Fig. 2a) and CITE-seq (Fig. 2b) experiments using the same set of antibodies on aliquots of the same pool of peripheral blood mononuclear cells. Using ADT levels we were able to construct cytometry-like ‘bi-axial’ gating plots (Fig. 2b) and compare these qualitatively and quantitatively to the flow cytometry data (Fig. 2a). Cell distribution profiles based on expression of marker proteins associated with various T cell subsets, B cells, plasmacytoid, myeloid dendritic cells and monocytes were remarkably similar (Fig. 2a,b; Supplemental Fig. 3a,b).

Next, we asked whether relative quantitative differences in expression levels observed by flow cytometry can be observed by CITE-seq. For this, we focused on the marker CD8a, since it exhibits a wide quantitative range of levels across immune cell populations. We incubated cord blood mononuclear cells (CBMCs) with CITE-seq antibody conjugates and fluorophore-conjugated antibodies, so that some CD8a epitopes on each cell would be labeled by fluorophore and some by oligo. Cells were sorted (by fluorescence-activated cell sorting, FACS) into separate pools based on CD8a fluorescence (CD8a very high (+++), CD8a high (++), CD8a intermediate(+) and CD8a low(+/−); Fig. 2c,d, Supplementary Fig. 3c). Each pool was then split and separately reanalyzed by flow cytometry and CITE-seq. For each pool defined by FACS, similar relative expression levels of CD8a were observed by both methods (Fig. 2e,f; Suppementary Fig. 3d,e). We conclude that CITE-seq ADT levels are consistent with gold-standard flow cytometry and can therefore enable high-resolution immunophenotyping in concert with transcriptomics.

We next aimed to perform a broad immunophenotypic and transcriptomic characterization of a complex immune cell population using CITE-seq. The immune system has been extensively profiled by cell surface markers[16] and scRNAseq[3,6,18], with both methods reliably identifying the same cell types at appropriate proportions. It is therefore an ideal system to validate the multimodal readout of CITE-seq. We prepared a CITE-seq panel of 13 well-characterized monoclonal antibodies that recognize cell-surface proteins routinely used as markers for immune-cell classification (Supplementary Table 2). Measuring protein abundance by antibodies can be complicated by non-specific binding, resulting in higher background and/or false positive results. To estimate background of antibody binding within experiments, we developed a low-level ‘spike-in’ control. We reasoned that a rare ‘spiked-in' population of murine cells could be easily distinguished transcriptomically but should not cross-react with our anti-human antibodies, enabling us to define background ADT levels directly from the data. We therefore spiked a low number of mouse fibroblasts (~4% 3T3) into our CBMCs, incubated the cell pool with our CITE-seq antibody panel and ran the 10× Genomics single cell workflow to measure cDNA and ADT profiles from 8,005 cells in total. Unsupervised graph-based clustering based on RNA expression revealed recognizable cell types indicated by expression of select marker genes (Fig. 3a, Supplementary Fig. 4). Murine cells clustered separately, and also exhibited low ADT counts for each marker, allowing us to set a baseline for signal vs noise to more clearly delineate positive from negative cell populations (Supplementary Fig. 5a,b). Through this thresholding step, we identified three antibody-oligo conjugates with no specific binding (*i.e.*, no signal over background threshold) and excluded these from further analysis (Supplementary Fig. 5b).

We detected strong ADT enrichment for different markers in the correct immune populations: We observe CD3e within the T cell cluster (Fig. 3b), and CD4 and CD8a in largely non-overlapping T cell subpopulations (Fig. 3b). We observe CD19 ADTs almost exclusively in the B cell cluster (Fig. 3b), CD56 and CD8a ADTs in the NK cluster, and CD11c, CD14 in the monocyte and dendritic cell cluster, together with CD16, also observed in the NK cells (Fig. 3b). We could also correctly identify a rare precursor cell population present at less than 2% in cord blood (CD34+ cells; Fig 3b). We observed that the ADT levels per-cell exhibited higher counts than mRNA-levels of the same genes and were less prone to 'drop-out' events. Consistent with this, we find low correlations between mRNA and ADT on a single cell basis and higher correlation when averaging expression within clusters (Supplementary Fig. 6). We used the ADT levels and clustering information based on transcriptome to construct multimodal CITE-seq 'bi-axial' gating plots, revealing similar profiles that are well-established by flow cytometry (Fig. 3c, Supplementary Fig. 5c). For example, we could resolve strong anti-correlation of CD4 and CD8a ADT levels in T cells and quantitative differences in marker expression between subsets, such as expression differences of CD8a between NK and T cells (blue and red cells; Fig. 3c) or CD4 between monocytes and T cells (yellow and turquoise cells; Fig. 3c). In addition to clustering cells based on their transcriptome, we performed clustering based on ADT levels, resulting in clear and consistent cell type separation (Supplementary Fig. 7).

We next asked whether CITE-seq could enhance our characterization of immune cell phenotypes, compared to scRNA-seq data alone. We noted an opposing gradient of CD56 and CD16 ADT levels within our transcriptomically-derived NK cell cluster, potentially corresponding to CD56^{bright} and CD56^{dim} subsets[19,20] (Fig. 3b, Supplementary Fig. 6a), and therefore sub-divided our NK cell cluster based on CD56 ADT levels (Fig. 3d). When comparing the molecular profiles of these groups, we observed protein and RNA changes that were highly consistent with literature[19,20]. We observed an apparent complementarity between levels of CD16 (Fig. 3e), and to a lesser extent of CD8a ADTs (Supplementary Fig. 6b), compared to CD56 ADTs within these two subsets. For 11 genes that have been previously characterized as differentially expressed within these subtypes[19–21], we detected up or downregulation consistent with the literature in 10 cases, including GZMB, GZMK and PRF1. This illustrates the potential for integrated and multimodal analyses to enhance discovery and description of cellular phenotypes, particularly when differentiating between cell populations with subtle transcriptomic differences.

The ability to layer additional molecular measurements on top of scRNA-seq data represents an exciting advance and growing challenge for the single cell community. CITE-seq enables multimodal analysis of single cells at the scale afforded by droplet-based single-cell sequencing approaches. We demonstrate the value of multimodal analysis to reveal phenotypes that could not be discovered by using scRNA-seq alone, and also envision CITE-seq enabling new studies of post-transcriptional gene regulation at the single cell level. In contrast to flow and mass cytometry, detection of oligo-barcoded antibodies is not limited by signal collision; a 10 nucleotide sequence can easily encode more barcodes than there are human proteins, enabling the potential for large-scale immunophenotyping with panels of tens to hundreds of antibodies. In addition, we envision that mild cell permeabilization and fixation procedures that are used for intracellular cytometry assays will also be compatible with CITE-seq, which may significantly expand the number of markers and biological questions that can be interrogated. A modified version of CITE-seq in which only ADTs are analyzed on a massively parallel scale without capturing cellular mRNAs (cytometry by sequencing) can also be envisaged. A conceptually similar approach, Abseq, has recently been described[22] which, in contrast to CITE-seq, focuses on the detection of single cell protein levels using DNA barcodes using highly advanced custom microfluidics.

Finally, we have shown that the CITE-Seq is fully compatible with a commercially-available single-cell instrument (10× Genomics) and should be readily adaptable to other droplet, microwell and combinatorial indexing-based high-throughput single-cell sequencing technologies[2,7–10] 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.

ONLINE METHODS

Conjugation of Antibodies to DNA-barcoding oligonucleotides

Highly specific, flow cytometry tested monoclonal antibodies (see below) were conjugated to oligonucleotides containing unique antibody identifier sequences and a polyA tail. We adopted a commonly used streptavidin-biotin interaction to link oligos to antibodies[23]. Antibodies were streptavidin labelled using the LYNX Rapid Streptavidin Antibody Conjugation Kit (Bio-Rad, USA), according to manufacturer’s instructions with modifications. Specifically, we labeled 15 µg of antibody with 10 µg of streptavidin. At this ratio, an average of two streptavidin tetramers will be conjugated per antibody molecule, which results in 8 binding sites for biotin on each antibody, on average. DNA oligonucleotides with a 5’ amine modification were purchased at IDT (USA) 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-labeled antibodies were incubated with biotinylated oligonucleotides in equimolar ratio (assuming two streptavidin tetramers per antibody on average) 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 50KDa MW cutoff (Millipore, USA). Removal of excess oligo was verified by 4% agarose gel electrophoresis (Supplementary Fig. 1a). Antibody-oligo conjugates were stored in PBS supplemented with sodium azide (0.05%) and BSA (1 μg/μl) at 4°C. See supplementary protocol for a more detailed description.

List of Antibodies used for CITE-seq

See Supplementary Table 2 for list antibodies, clones and barcodes used for CITE-seq.

Antibody-oligo sequences

We leverage the DNA-dependent DNA polymerase activity of commonly used reverse transcriptases[24] to convert CITE-seq DNA oligonucleotides into cDNA during reverse transcription together with mRNAs. The DNA-dependent DNA polymerase activity of MMLV reverse transcriptases is well established. All SMART (Switching Mechanism at 5' end of RNA Template) library prep protocols (e.g. commercialized by Clontech) rely on this activity: The RT enzyme switches at the end of the RNA template to a template switch oligo (TSO), which is mainly DNA, for further cDNA synthesis. This activity has been shown to be highly sensitive and reproducible. Single cell RNA-seq protocols (including 10× Genomics and Drop-seq) also entirely rely on this activity to append a PCR handle to the 5'end of full length cDNAs which is used for subsequent amplification. 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.* 10× Single Cell 3' v1 uses read 1 while Drop-seq and 10× Single Cell 3' v2 use read 2). Our proof-of-principle human/mouse antibody-barcoding oligonucleotide designs included UMIs, which are redundant for Drop-seq and 10× protocols due to the UMI addition to the cDNA at reverse transcription. UMIs on the antibody-conjugated oligonucleotide 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/
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAATNNBAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAA

BC12: /5AmMC12/
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGTANNBAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAA

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

BC6: /5AmMC12/
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCAATNNBAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAA

BC12: /5AmMC12/
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGTANNBAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAA

CBMC profiling – (Drop-seq and 10× v2 compatible oligos, containing TruSeq small RNA read 2 handle)

v2_BC1: /5AmMC12/
CCTTGGCACCCGAGAATTCCAATCACGBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC2: /5AmMC12/
CCTTGGCACCCGAGAATTCCACGATGTBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC3: /5AmMC12/
CCTTGGCACCCGAGAATTCCATTAGGCBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC4: /5AmMC12/
CCTTGGCACCCGAGAATTCCATGACCABAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC6: /5AmMC12/
CCTTGGCACCCGAGAATTCCAGCCAATBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC9: /5AmMC12/
CCTTGGCACCCGAGAATTCCAGATCAGBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC10: /5AmMC12/
CCTTGGCACCCGAGAATTCCATAGCTTBAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAA

v2_BC12: /5AmMC12/
CCTTGGCACCCGAGAATTCCACTTGTABAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC8: /5AmMC12/
CCTTGGCACCCGAGAATTCCAACTTGABAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC11: /5AmMC12/
CCTTGGCACCCGAGAATTCCAGGCTACBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC13: /5AmMC12/
CCTTGGCACCCGAGAATTCCAAGTCAABAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC14: /5AmMC12/
CCTTGGCACCCGAGAATTCCAAGTTCCBAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

v2_BC5: /5AmMC12/
CCTTGGCACCCGAGAATTCCAACAGTGBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Cell ‘staining’ with DNA-barcoded antibodies for CITE-seq

Roughly 500,000 cells were resuspended in cold PBS containing 2% BSA and 0.01% Tween and filtered through 40 μm 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. Antibody concentrations were 1 μg per test as recommended by the manufacturer (Biolegend, USA) for flow cytometry applications. Cells were washed 3× by resuspension in PBS containing 2% BSA and 0.01% Tween, followed by centrifugation (~480*g* 5 minutes at 4°C) and supernatant exchange. After the final wash cells were resuspended at appropriate cell concentration in PBS for Drop-seq[1] or 10× Genomics[3] applications. See supplementary protocol for a more detailed description.

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 mRNA were converted into sequencing libraries by tagmentation as described[1]. After quantification, libraries were merged at appropriate concentrations (10% of a lane for ADT, 90% cDNA library). Sequencing was performed on a HiSeq 2500 Rapid Run with v2 chemistry per manufacturer’s instructions (Illumina, USA). See supplementary protocol for a more detailed description.

10× – CITE-seq

The 10× single cell run was performed according to the manufacturer’s instructions (10× Genomics, USA) with modifications. For the Human/Mouse mixing experiment (ran on Single Cell 3’ version 1) ~17,000 cells were loaded to yield around ~10,000 cells with an intermediate/high doublet rate. For CBMC profiling (ran on Single Cell 3’ version 2), ~7,000 cells were loaded to obtain a yield of ~4,000 cells. For CBMC profiling we spiked-in mouse cells at low frequency (~4%). This allowed us to draw antibody signal-to-noise cutoffs and allowed us to estimate the true doublet rates (4%) in our experiments, and compare these to the estimates provided by the equipment manufacturer (~3.1%) (see below). 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 for the Single Cell 3' v1 experiments, or according to manufacturer's instructions for the Single Cell 3' v2 experiments. ADT and cDNA libraries were merged and sequenced as described above. See supplementary protocol for a more detailed description.

Cell culture

HeLa (human), 4T1 (mouse) and 3T3 (mouse) cells were 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, HeLa and 4T1 cells were mixed in equal proportions and incubated with DNA barcoded CITE-seq antibodies as described above. For the low frequency mouse spike-ins ~5% 3T3 cells were mixed into CBMC pool before performing CITE-seq.

Blood mononuclear cells

Cord blood mononuclear cells (CBMCs) were isolated from cord blood (New York Blood Center) as described[25]. Cells were kept on ice during and after isolation. Peripheral blood mononuclear cells were obtained from Allcells (USA).

Comparing flow cytometry and CITE-seq

Cells were stained with a mixture of fluorophore (CD8a-FITC, Biolegend, USA) labelled antibodies and CITE-seq oligo labelled antibodies from the same monoclonal antibody clone (RPA-T8) targeting CD8a, at concentrations recommended by the manufacturer (1ug per test, Biolegend, USA). Cells were also stained with Anti-CD4-APC antibody (RPA-T4, Biolegend, USA). Cells were sorted into pools of different CD8a expression levels using the Sony SH800 cell sorter, operated per manufacturer's instructions. Pools were then split into two and reanalyzed by flow cytometry using Sony SH800 or processed for CITE-seq using Drop-seq as described above. Flow cytometry data was plotted using FlowJo v9 (USA).

Multiparameter flow cytometry

Cells were stained with the following mouse anti-human antibodies, purchased from BD Biosciences (USA): CD3e Hilyte 750 Allophycocyanin (H7APC), CD4 Brilliant Blue (BB) 630, CD8a Phycoerythrin (PE), CD14 Brilliant Violet (BV) 750, CD19 BV570, CD11c Cyanin5 PE, CD2 Brilliant Ultraviolet (BUV) 805, and CD57 BB790. After washing cells in PBS and fixing in 0.5% paraformaldehyde, samples were acquired on a BD Symphony A5 flow cytometer and data was analyzed using FlowJo v9 (USA).

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). 10× data from the species mixing experiment was processed using Cell Ranger 1.2 using default parameters and no further filtering was applied. 10× data from CBMC experiments (v2 chemistry) was processed using the same pipeline as Drop-seq data. Reads were aligned to the human reference sequence GRCh37/hg19 (CD8a FACS comparison), or an hg19 and mouse reference mm10 concatenation (species mixing experiment, CBMCs). 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 CD8a FACS comparison data, 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 200 genes detected and cells with a total number of UMIs or genes (in log10 after adding a pseudo-count) that is more than 3 standard deviations above or below the mean. The same filtering strategy was used for the CBMC data, the only difference being a gene threshold of 500.

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 CBMCs 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 passed the RNA-specific filters and had a minimum number of total ADT counts (species mixing: 10, CD8a FACS comparison: 1, CBMC: 50).

CBMC RNA normalization and clustering—After read-alignment and cell filtering, we assigned the species to each cell barcode. If more than 90% of UMI counts were coming from human genes, the cell barcode was considered to be human. If it was less than 10% the assigned species was mouse. Cell barcodes in between were considered mixed species. The resulting assignment was human: 8005, mouse: 579, mixed: 33. Unless stated otherwise, analysis was performed on only the human cells and genes from the human reference genome.

We converted the matrix of UMI counts into a log-normalized expression matrix x with $x_{i,j}=\log(\frac{c_{i,j}*10,000}{m_j})$, 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 556 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 1 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 5% constituted our significant number of dimensions (in this case the number was 13). For clustering we used a modularity optimization algorithm that finds community structure in the data[26]. The data is represented as a weighted network with cells being nodes and squared

Jaccard similarities as edge weights (based on Euclidian distance of significant PCs and a neighborhood size of 40 (0.5% of all cells)). The clustering algorithm, as implemented in the “cluster_louvain” function of the igraph R package, find a partitioning of the cells with high density within communities as compared to between communities. For 2D visualization we further reduced the dimensionality of the data to 2 using t-SNE[27,28].

CBMC 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 13 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 $y=\mathrm{clr}(x)=[\ln\left(\frac{x_1}{g(x)}\right),\ln\left(\frac{x_2}{g(x)}\right),\dots,\ln\left(\frac{x_5}{g(x)}\right)$ 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 separately for human and mouse cells (Supplementary Fig. 5a,b). For each ADT we determined the mean and variance of the mouse cells and defined the species-independent cutoff (separating ‘off’ state from ‘on’ state where protein is present) to be one standard deviation larger than the mean.

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 we subtracted the mouse-derived cutoffs from the CLR-transformed ADT counts for each antibody. Cell-to-cell weights were squared Jaccard similarities based on Euclidean distance and neighborhood size of 0.5% of the total number of cells.

Estimation of doublet rate using low frequency mouse spike-in—Spiking-in mouse cells at low frequency allowed us to estimate the true doublet rates (4%) in our CMBC profiling experiment, and compare these to the estimates provided by the equipment manufacturer (~3.1%). For estimation of the doublet rate in our experiments we modeled the droplet cell capture process as a Poisson distribution with a loading rate lambda and a fixed mouse fraction of 6.5%. We optimized lambda so that simulated data would most closely match the observed species distribution. The resulting lambda was 0.068 and the doublet rate (fraction of droplets with more than one cell of all droplets with at least one cell) observed in the simulations was 4%.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge members of the NYGC Technology Innovation lab S. Jaini and K. Pandit for critical discussions and support. We thank E. Papalexi for help with CBMCs isolation. We thank M. Coppo, S. Fennessey, B. Baysa and S. Pescatore at NYGC for sequencing support. We thank C. Kocks for discussions of the manuscript.

Multiparameter flow cytometer instrument time and reagents were generously provided by the Vaccine Research Center of the National Institutes of Health. CH was supported by a Deutsche Forschungsgemeinschaft research fellowship. Research reported in this publication was partially supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number UM1HG008901, and an NIH New Innovator Award (DP2-HG-009623) to RS.

References

1. Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015; 161:1202–1214. [PubMed: 26000488]
2. Klein AM, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015; 161:1187–1201. [PubMed: 26000487]
3. Zheng GXY, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8:1–12.
4. Pontén F, et al. A global view of protein expression in human cells, tissues, and organs. Mol Syst Biol. 2009; 5:1–9.
5. Paul F, et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell. 2015; 163:1663–1677. [PubMed: 26627738]
6. Wilson NK, et al. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell. 2015; 16:712–724. [PubMed: 26004780]
7. Yuan J, Sims PA. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. Scientific Reports. 2016; 6:33883. [PubMed: 27670648]
8. Gierahn TM, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods. 2017; 14:395–398. [PubMed: 28192419]
9. Cao J, et al. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. Biorxiv. 2017; :1–35. DOI: 10.1101/104844
10. Rosenberg AB, et al. Scaling single cell transcriptomics through split pool barcoding. Biorxiv. 2017; :1–13. DOI: 10.1101/105163
11. Ståhlberg A, Thomsen C, Ruff D, Åman P. Quantitative PCR analysis of DNA, RNAs, and proteins in the same single cell. Clinical Chemistry. 2012; 58:1682–1691. [PubMed: 23014600]
12. Genshaft AS, et al. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. Genome Biology. 2016; 17:188. [PubMed: 27640647]
13. Albayrak C, et al. Digital Quantification of Proteins and mRNA in Single Mammalian Cells. Molecular Cell. 2016; 61:914–924. [PubMed: 26990994]
14. Darmanis S, et al. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. Cell Reports. 2016; 14:380–389. [PubMed: 26748716]
15. Frei AP, et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nature Methods. 2016; 13:269–275. [PubMed: 26808670]
16. Murphy, K., Travers, P., Walport, M. Janeways Immuno Biology 7th ed. Garland Pub. Inc; New York and London: 2008.
17. Robinson JP, Roederer M. History of Science. Flow cytometry strikes gold. Science. 2015; 350:739–740. [PubMed: 26564833]
18. Fan HC, Fu GK, Fodor SPA. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015; 347:1258367–1258367. [PubMed: 25657253]
19. Poli A, et al. CD56bright natural killer (NK) cells: an important NK cell subset. Immunology. 2009; 126:458–465. [PubMed: 19278419]
20. Ferlazzo G, Munz C. NK Cell Compartments and Their Activation by Dendritic Cells. The Journal of Immunology. 2004; 172:1333–1339. [PubMed: 14734707]
21. Wendt K, et al. Gene and protein characteristics reflect functional diversity of CD56dim and CD56bright NK cells. Journal of Leukocyte Biology. 2006; 80:1529–1541. [PubMed: 16966385]
22. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Scientific Reports. 2017; 7:44447. [PubMed: 28290550]

23. Adler M, Wacker R, Niemeyer CM. Sensitivity by combination: immuno-PCR and related technologies. Analyst. 2008; 133:702–18. [PubMed: 18493669]
24. Baranauskas A, et al. Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. Protein Eng. Des. Sel. 2012; 25:657–668. [PubMed: 22691702]
25. Breton GEL, Lee J, Liu K, Nussenzweig MC. Defining human dendritic cell progenitors by multiparametric flow cytometry. Nature Protocols. 2015; 10:1407–1422. [PubMed: 26292072]
26. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008; 2008:P10008.
27. Maaten LVD, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008; 9:2579–2605.
28. van der Maaten L. Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research. 2014

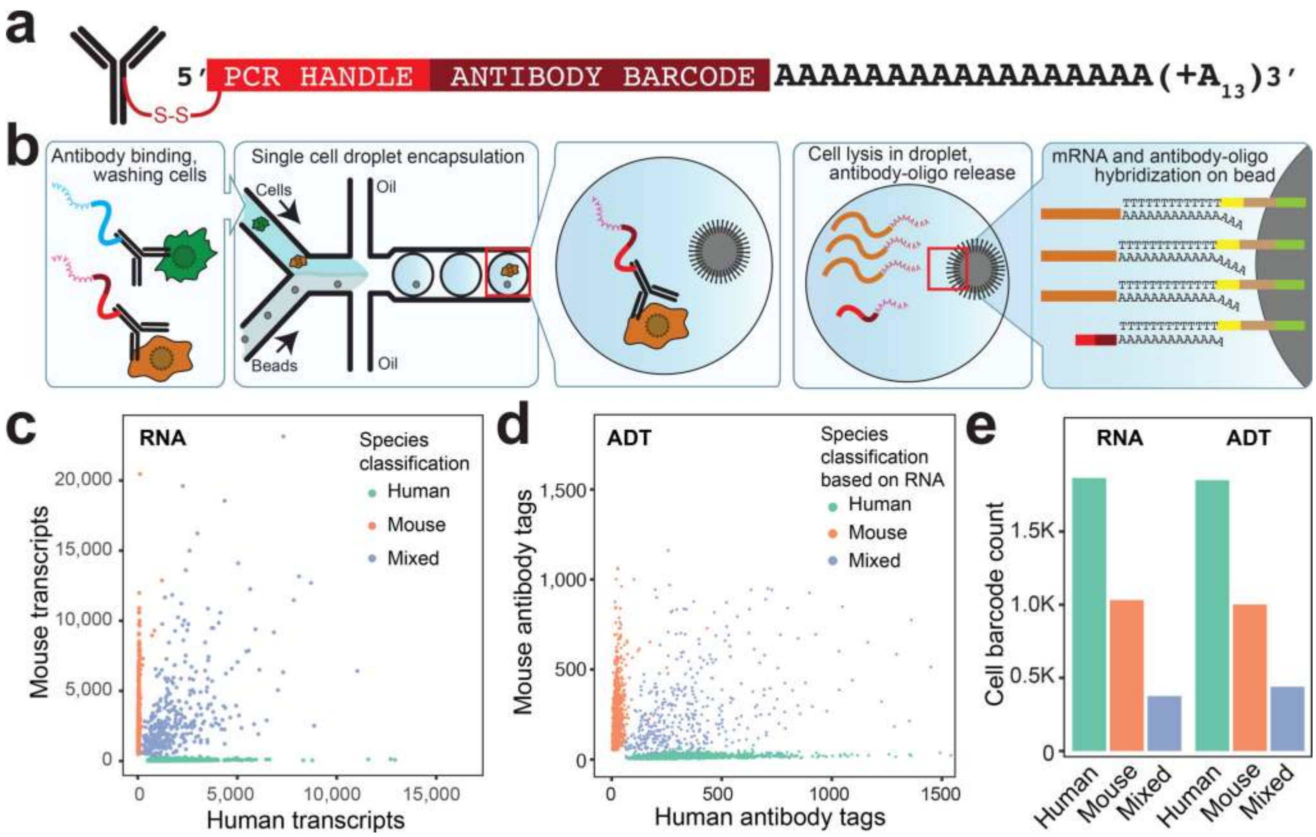

Figure 1. CITE-seq enables simultaneous detection of single cell transcriptomes and protein markers

(**a**) Illustration of the DNA-barcoded antibodies used in CITE-seq. (**b**) Schematic representation of CITE-seq in combination with Drop-seq[1]. Cells are incubated with antibodies, washed and passed through a microfluidic chip where a single cell and one bead are occasionally encapsulated in the same droplet. After cell lysis mRNAs and antibody-oligos anneal to oligos on Drop-seq beads, linking cell barcodes with cellular transcripts and antibody-derived oligos. (**c – e**) 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 processed by Drop-seq. (**c**) Quantification of the number of human and mouse transcripts associating to each cell barcode. Green: >90% human reads, Red: >90% mouse reads, Blue: >10% human and mouse (multiplet). (**d**) Quantification of antibody tags (ADTs) associated with each cell barcode. Points are colored based on species classifications using transcripts in (c). (**e**) Quantification of human, mouse or mixed-cell barcodes based on RNA transcripts, or ADTs.

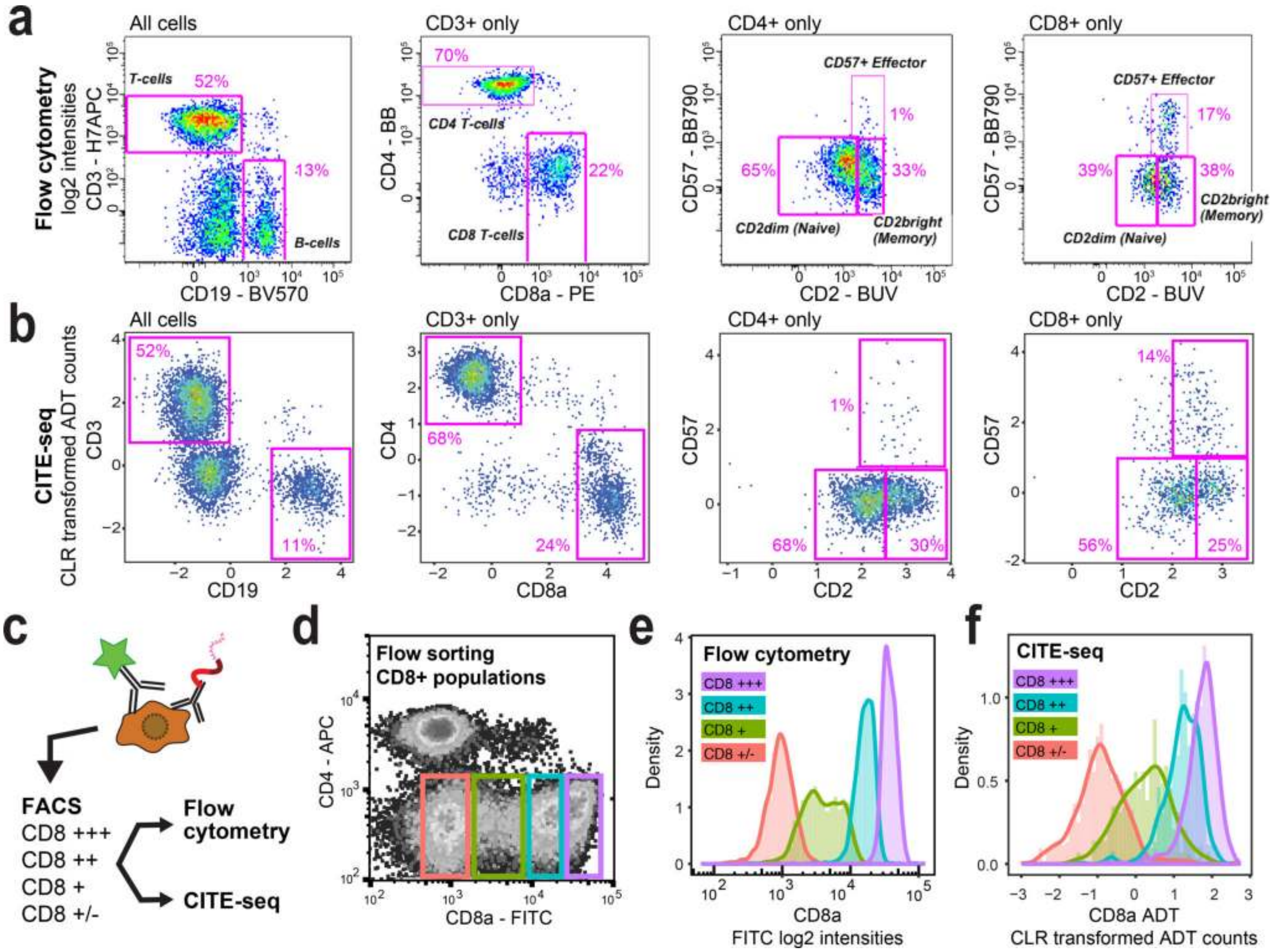

Figure 2. Qualitative and quantitative comparison between CITE-seq and flow cytometry

(a–b) Comparison of qualitative readout of flow cytometry to CITE-seq. Aliquots of cells from the same pool were processed for flow cytometry (a) and CITE-seq (b). Functionally important immune subsets were selected based on their established flow cytometry expression patterns and their relative frequencies compared to the entire population, and within the CD3e, CD4 and CD8a positive subsets. **(c)** Illustration of experiment for relative quantitative comparison of flow cytometry and CITE-seq. (**d**) Profile of CD4 and CD8a fluorescence in CBMCs. Colored boxes are gates set to sort cells with different levels of CD8a expression. (**e**) Flow cytometry of cells sorted in panel d. Merged histograms of CD8a levels in the four different pools of cells. (**f**) CD8a levels of the different pools of cells sorted in panel d, as measured by CITE-seq. Merged histograms of four CITE-seq runs of the separate pools.

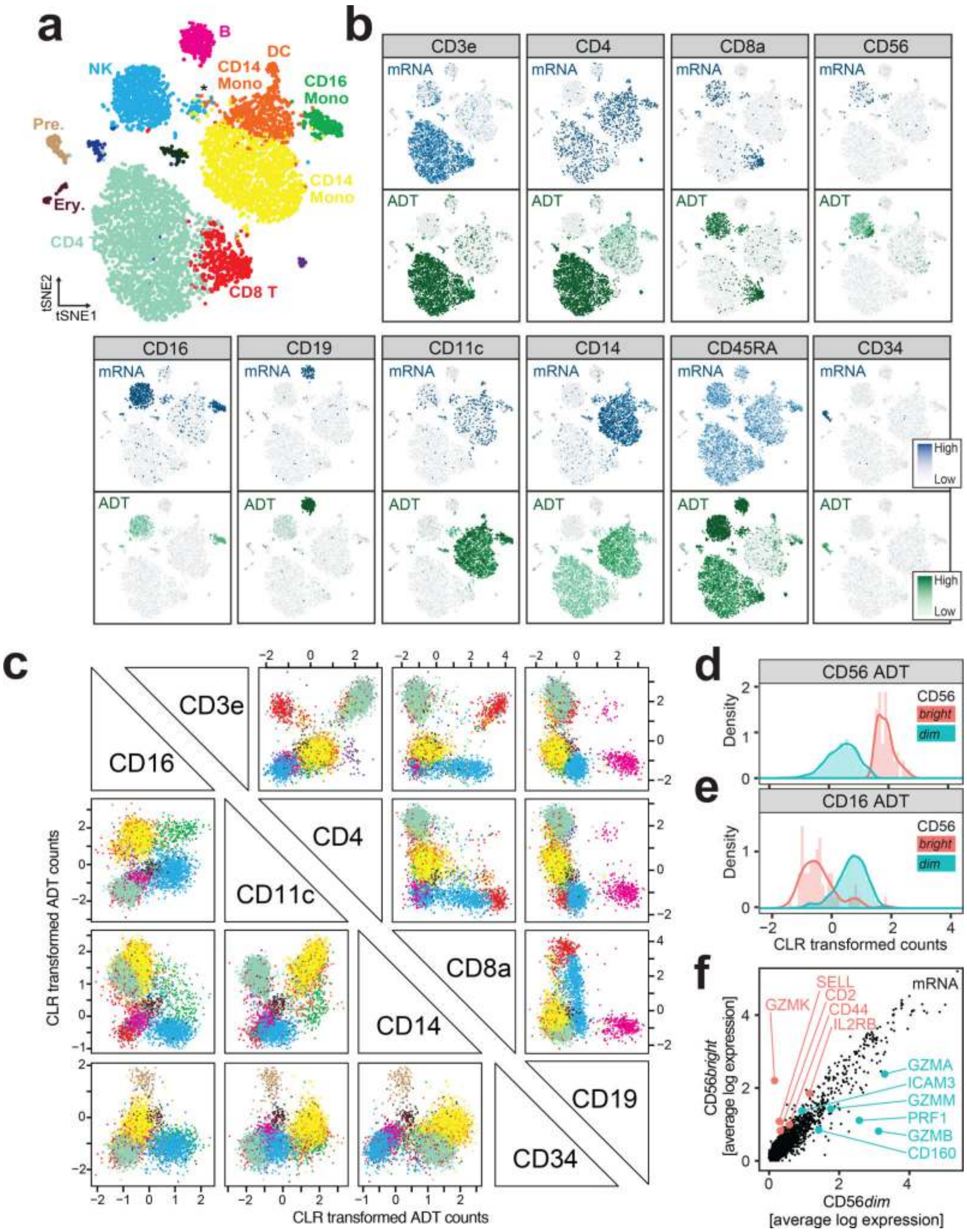

Figure 3. CITE-seq allows detailed multimodal characterization of cord blood mononuclear cells

(**a**) Clustering of 8,005 CITE-seq single-cell expression profiles of CBMCs reveals distinct cell populations based on transcriptome. The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among all cells. Major cell types in cord blood can be discerned based on marker gene expression (Supplementary Fig. 4). Putative doublets co-expressing multiple lineage markers (*) are indicated. The mouse control cell population was excluded from the clustering. (**b**) mRNA (blue) and corresponding ADT (green) signal for the CITE-seq antibody panel projected on the tSNE plot from panel a. Darker shading corresponds to higher levels measured. (**c**) Multimodal bi-axial plots.

Pairwise comparison of different ADT levels in single cells for select markers (see Supplementary Fig. 5c for all markers). ADT counts were centered log-ratio transformed and plotted with colors based on RNA clusters shown in panel a. (**d–f**) NK cells were split into CD56^{bright} and CD56^{dim} groups based on CD56 ADT levels. Histogram of CD56 (d) and CD16 (e) levels in the CD56^{bright} and CD56^{dim} groups. (**f**) Differential gene expression analysis between the CD56^{bright} and CD56^{dim} cells. Genes known from literature to be higher expressed in CD56^{bright} are marked in red, genes known to be higher in CD56^{dim} are marked in green.