© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Simultaneous epitope and transcriptome measurement in single cells

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

High-throughput single-cell RNA sequencing has transformed our understanding of complex cell populations, but it does not provide phenotypic information such as cell-surface protein levels. Here, we describe cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), a method in which oligonucleotide-labeled antibodies are used to integrate cellular protein and transcriptome measurements into an efficient, single-cell readout. CITE-seq is compatible with existing single-cell sequencing approaches and scales readily with throughput increases.

The unbiased and high-throughput nature of modern single-cell RNA-seq (scRNA-seq) approaches has proven invaluable for describing heterogeneous cell populations[1–3]. Prior to single-cell genomics, cellular states were routinely described using 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 from a cell sorter with single-cell transcriptomics; this process allows immunophenotypes to be mapped onto transcriptomically derived clusters. However, massively parallel approaches based on droplet microfluidics[1–3], microwells[7,8] or combinatorial indexing[9,10] are incompatible with cytometry and therefore cannot be augmented with protein information. Targeted methods to simultaneously measure transcripts and proteins in single cells are limited in scale or can only profile a few genes and proteins in parallel[11–15] (**Supplementary Table 1**).

Here, we describe CITE-seq, a method that combines highly multiplexed protein marker detection with unbiased transcriptome profiling for thousands of single cells. We demonstrate that the method is readily adaptable to two high-throughput scRNA-seq applications and show that multimodal data analysis can achieve a more detailed characterization of cellular phenotypes than transcriptome measurements alone.

We devised a digital, sequencing-based readout for protein levels by conjugating antibodies to oligonucleotides (oligos) that can be captured by oligo-dT primers (used in most scRNA-seq library preparations), contain a barcode for antibody identification and include a handle for PCR amplification (see Online Methods). A commonly used streptavidin–biotin interaction links the 5′ end of oligos to antibodies, and a disulfide bond allows the oligo to be released in reducing conditions (**Fig. 1a** and **Supplementary Fig. 1a**). The antibody–oligo complexes are incubated with single-cell suspensions in conditions comparable to flow cytometry staining protocols; after this incubation, cells are washed to remove unbound antibodies and processed for scRNA-seq. In our example, we encapsulated single cells into nanoliter-sized aqueous droplets in a microfluidic apparatus designed to perform Drop-seq[1] (**Supplementary Fig. 1b**). After cell lysis in droplets, cellular mRNAs and antibody-derived oligos both anneal via their 3′ polyA tails to Drop-seq beads containing oligo-dT (**Supplementary Fig. 1b**,**c**) and are indexed by a shared cellular barcode during reverse transcription. The amplified cDNAs and antibody-derived tags (ADTs) can be separated by size and converted into Illumina-sequencing libraries independently (**Supplementary Fig. 1d**). Importantly, because the two library types are generated separately, their relative proportions can be adjusted in a pooled single lane to ensure that the required sequencing depth is obtained for each library.

To assess our method's ability to distinguish single cells based on surface protein expression, we designed a proof-of-principle 'species-mixing' experiment that leverages the species-specific and highly expressed marker CD29 (Integrin beta-1). A 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. 1** and **Supplementary Fig. 2a**,**b**). 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. Most droplets (97.2%) that were identified as containing human, mouse or mixed cells by transcriptome (**Fig. 1b**) received the same species classification by ADT counts (**Fig. 1c**). Cell counts based on RNA or ADT are highly correlated between both methods (**Supplementary Fig. 2b**), and this demonstrates the low dropout rate of ADT signals. We performed the same experiment using a commercially available system from 10x Genomics and obtained comparable results (**Supplementary Fig. 2c–f**).

We sought to characterize the quantitative nature of the CITE-seq protein readout. Flow cytometry is the gold standard for

[1]New York Genome Center, New York, New York, USA. [2]New York University Medical Center, New York, New York, USA. [3]New York University Center for Genomics and Systems Biology, New York, New York, USA. Correspondence should be addressed to M.S. (mstoeckius@nygenome.org).

RECEIVED 2 MARCH; ACCEPTED 7 JULY; PUBLISHED ONLINE 31 JULY 2017; DOI:10.1038/NMETH.4380

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

enumerating 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 CITE-seq antibodies directed against common flow cytometry markers to identify and discriminate immune subpopulations. 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 'biaxial' 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** and **Supplementary Fig. 3a**,**b**).

Next, we asked whether quantitative differences in expression observed by flow cytometry can be observed by CITE-seq. For this, we focused on the marker CD8a, since its levels vary widely 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 subjected to fluorescence-activated cell sorting (FACS) into separate pools based on CD8a fluorescence (very high (+++), high (++), intermediate (+) and low (+/−); **Fig. 2c**,**d** and **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 CD8a expression levels were observed by both methods (**Fig. 2e**,**f** and **Supplementary 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.

The immune system has been extensively profiled using cell-surface markers[16] and scRNAseq[3,6,18], and both methods reliably identify the same cell types at consistent proportions. A complex immune cell population is therefore an ideal system for validating 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**). To estimate nonspecific background antibody binding within experiments, we developed a low-level 'spike-in' control. A rare spiked-in population of murine cells should be easily distinguished transcriptomically but should not cross-react with our anti-human antibodies; this would enable us to define background ADT levels directly from the data. We therefore spiked mouse 3T3 fibroblasts (~4%) into our CBMCs, incubated the cell pool with our CITE-seq antibody panel and ran the 10x Genomics single-cell workflow on a total of 8,005 cells. Unsupervised graph-based clustering using RNA expression revealed recognizable cell types that express select marker genes (**Fig. 3a** and **Supplementary Fig. 4**). Murine cells clustered separately (data not shown) and exhibited low ADT counts for each marker, and this allowed us to set a baseline for signal versus 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**).

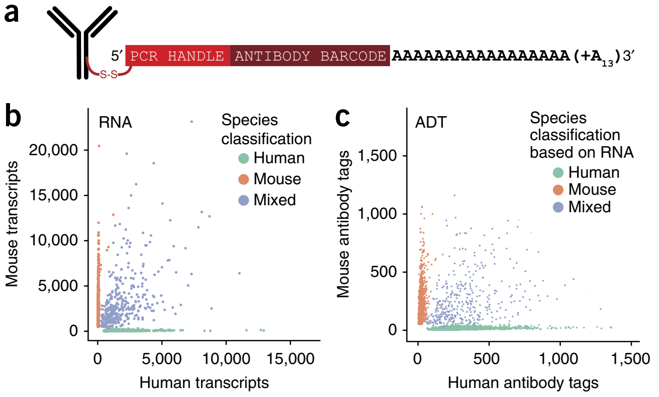

Figure 1 | CITE-seq enables simultaneous detection of single cell transcriptomes and protein markers. (**a**) Illustration of the DNA-barcoded antibodies used in CITE-seq. (**b**,**c**) Analysis of mixtures of mouse and human cells incubated with oligo-tagged-antibodies specific for either human or mouse CD29 (integrin beta) and processed by Drop-seq. (**b**) Number of transcripts associated with each cell barcode. Green, >90% human reads; red, >90% mouse reads; blue, >10% human and mouse (multiplet). (**c**) Number of antibody tags (ADTs) associated with each cell barcode. Points are colored based on species classifications using transcripts in **b**.

We detect strong ADT enrichment in the correct immune populations—CD3e within the T-cell cluster; CD4 and CD8a in largely nonoverlapping T-cell subpopulations; CD19 almost exclusively in B-cells; CD56, CD16 and CD8a in the NK cluster; and CD11c and CD14 in the monocyte and dendritic cell cluster (**Fig. 3b**). We can also identify a rare precursor cell population at less than 2% in cord blood (CD34+ cells; **Fig. 3b**). Per-cell ADT counts are higher than mRNA levels for the same genes and are less prone to 'dropout' 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 transcriptome-based clustering information to construct multimodal CITE-seq 'biaxial' gating plots; this revealed similar profiles that are well-established by flow cytometry (**Fig. 3c** and **Supplementary Fig. 5c**). For example, we could resolve strong anticorrelation of CD4 and CD8a ADT levels in T cells and quantitative differences in marker expression between subsets—these included expression differences of CD8a between NK and T cells (blue and red cells; **Fig. 3c**) or of CD4 between monocytes and T cells (yellow and turquoise cells; **Fig. 3c**). In addition, clustering based on ADT levels results in clear and consistent cell-type separation (**Supplementary Fig. 7**).

We next asked whether multimodal data from CITE-seq could enhance the characterization of immune cell phenotypes compared to scRNA-seq alone. We noted an opposing gradient of CD56 and CD16 ADT levels within our transcriptomically derived NK cell cluster, potentially revealing CD56$^{\text{bright}}$ and CD56$^{\text{dim}}$ subsets[19,20] (**Fig. 3b** and **Supplementary Fig. 8a**); therefore, we subdivided 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 the literature[19,20]. We observed an apparent complementarity between levels of CD16 (**Fig. 3e**)—and to a lesser extent of CD8a ADTs (**Supplementary Fig. 8b**)—compared with CD56 ADTs within these two subsets. For 11 genes

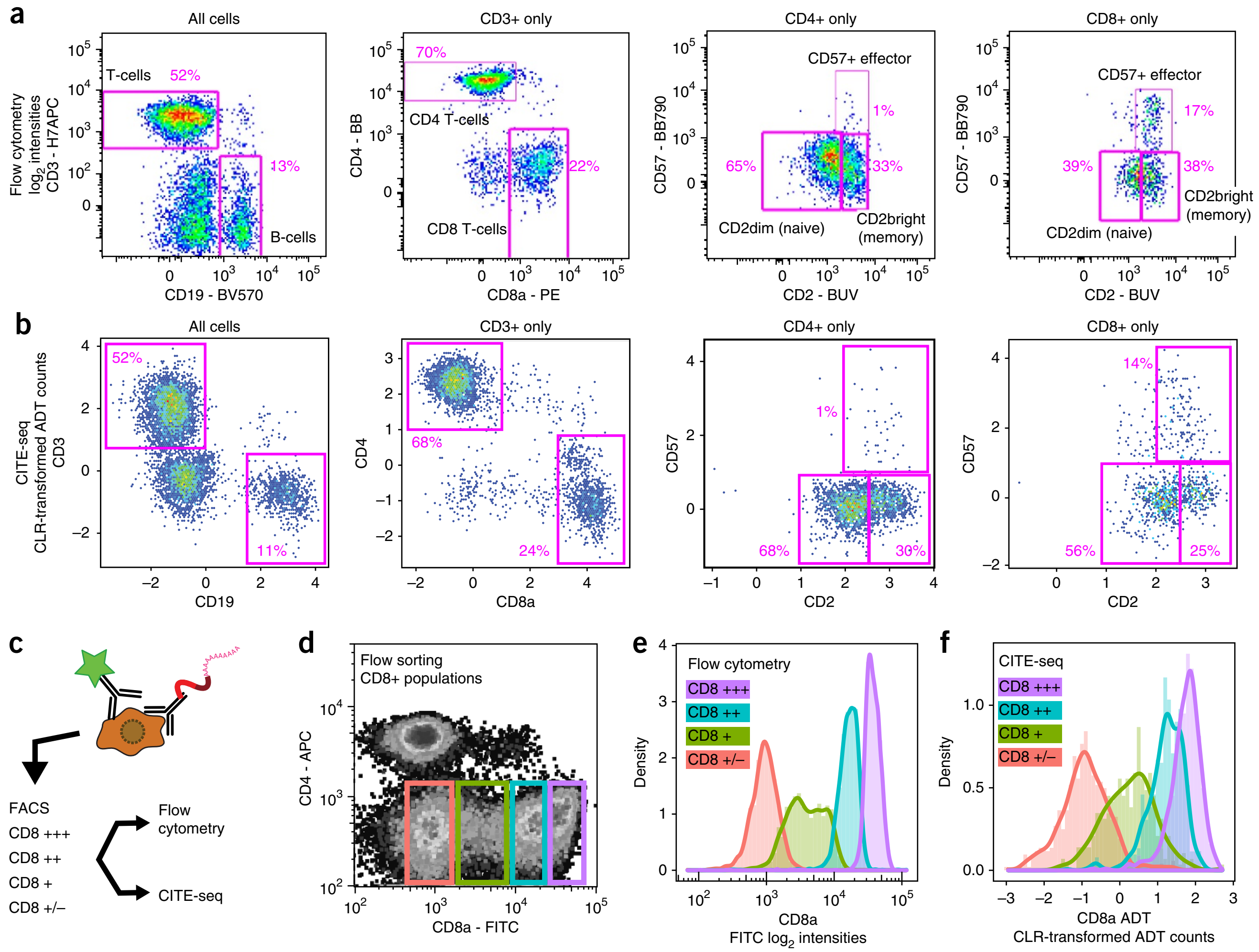

Figure 2 | CITE-seq is qualitatively and quantitatively similar to flow cytometry. (**a**–**b**) Comparison of qualitative readout of flow cytometry to CITE-seq. Aliquots of cells from the same pool were processed for flow cytometry (**a**) and CITE-seq (**b**). Functional immune subsets were selected based on established flow cytometry expression patterns and relative frequencies compared to the entire population or to the CD3e-, CD4- and CD8a-positive subsets. (**c**) Experimental design of relative quantitative comparison. (**d**) Profile of CD4 and CD8a fluorescence in CBMCs. Colored boxes indicate CD8a-expression-sorting gates. (**e**) Merged histograms of CD8a levels measured by flow cytometry in the four pools of cells sorted in **d**. (**f**) Merged histograms of CD8a levels measured by CITE-seq in the four pools of cells sorted in **d**. CLR, centered log ratio. H7APC, BV570, BB, PE, BUV, BB790, FITC, APC; flow cytometry fluorophores.

that have previously been characterized as differentially expressed within these subtypes[19–21], we detected upregulation or downregulation consistent with the literature in ten cases, including those of *GZMB*, *GZMK* and *PRF1* (**Fig. 3f**). 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 direction for the single-cell research 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 we also envision the use of CITE-seq for 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-nt sequence can easily encode more barcodes than there are human proteins, and this enables large-scale immunophenotyping with panels of tens to hundreds of antibodies. In addition, mild cell permeabilization and fixation procedures used for intracellular cytometry assays should also be compatible with CITE-seq, and they may significantly expand the number of useful markers. 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]; this approach, in contrast to CITE-seq, focuses on the detection of single-cell protein levels using DNA barcodes and highly advanced custom microfluidics.

Finally, we have shown that the CITE-seq is fully compatible with a commercially available single-cell platform (10x Genomics) and should be readily adaptable to other droplet-, microwell- and combinatorial-indexing-based high-throughput single-cell sequencing technologies[2,7–10] with either no or minor customizations.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

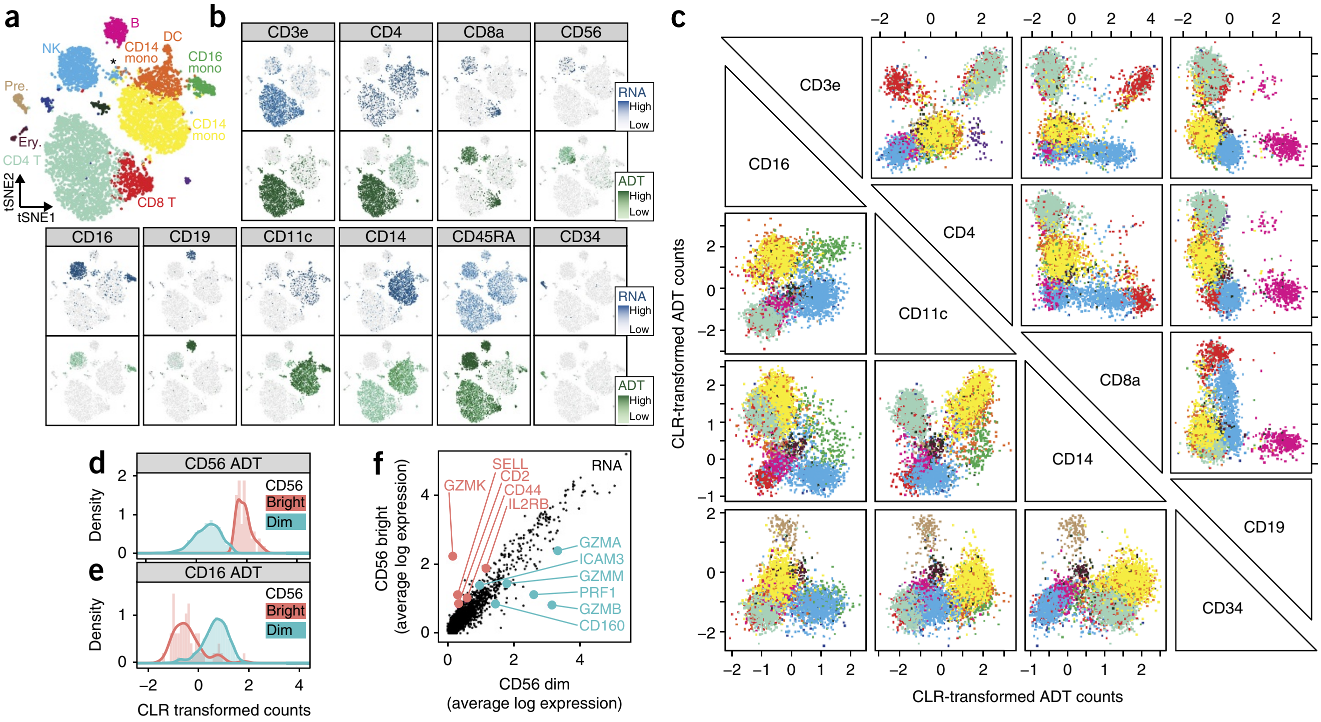

Figure 3 | CITE-seq allows detailed multimodal characterization of cord blood mononuclear cells. (**a**) Transcriptome-based clustering of 8,005 CITE-seq single-cell expression profiles of CBMCs reveals distinct cell populations. Major cord blood cell types can be discerned by marker gene expression (**Supplementary Fig. 4**). B, B cells; T, T cells; NK, natural killer cells; mono, monocytes; DC, dendritic cells; pre., precursors; ery., erythrocytes/blasts. Putative doublets coexpressing 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**. (**c**) Multimodal biaxial 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 in panel **a**. (**d**–**f**) NK cells are split into CD56^{bright} and CD56^{dim} populations based on CD56 ADT levels. Histogram of CD56 (**d**) and CD16 (**e**) levels in the CD56^{bright} and CD56^{dim} populations. (**f**) Differential gene expression analysis between the CD56^{bright} and CD56^{dim} cells. Genes known from literature to be expressed more highly in CD56^{bright} are marked in red; genes known to be expressed more highly in CD56^{dim} are marked in green.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge S. Jaini and K. Pandit of the NYGC Technology Innovation Lab for critical discussions and support; E. Papalexi for help with CBMCs isolation; M. Coppo, S. Fennessey, B. Baysa and S. Pescatore at NYGC for sequencing support; and C. Kocks for manuscript discussions. Multiparameter flow cytometer instrument time and reagents were generously provided by the Vaccine Research Center of the National Institutes of Health. C.H. was supported by a Deutsche Forschungsgemeinschaft research fellowship. R.S. was supported by an NIH New Innovator Award (DP2-HG-009623). Research reported here was partially supported by the National Human Genome Research Institute under Award Number UM1HG008901.

AUTHOR CONTRIBUTIONS

M.S. conceived and designed the study with input from B.H.-L., R.S., H.S. and P.S. M.S. performed all experiments. C.H. and R.S. designed and contributed the computational analyses. W.S. assisted with Drop-seq experiments. P.K.C. provided conceptual input on how to benchmark CITE-seq to flow cytometry and performed multiparameter flow cytometry analysis. M.S., C.H., R.S. and P.S. interpreted the data. M.S. and P.S. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
2. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
3. Zheng, G.X.Y. *et al. Nat. Commun.* **8**, 14049 (2017).
4. Pontén, F. *et al. Mol. Syst. Biol.* **5**, 337 (2009).
5. Paul, F. *et al. Cell* **163**, 1663–1677 (2015).
6. Wilson, N.K. *et al. Cell Stem Cell* **16**, 712–724 (2015).
7. Yuan, J. & Sims, P.A. *Sci. Rep.* **6**, 33883 (2016).
8. Gierahn, T.M. *et al. Nat. Methods* **14**, 395–398 (2017).
9. Cao, J. *et al.* Preprint at http://www.biorxiv.org/content/early/2017/02/02/104844 (2017).
10. Rosenberg, A.B. *et al.* Preprint at http://www.biorxiv.org/content/early/2017/02/02/105163 (2017).
11. Ståhlberg, A., Thomsen, C., Ruff, D. & Åman, P. *Clin. Chem.* **58**, 1682–1691 (2012).
12. Genshaft, A.S. *et al. Genome Biol.* **17**, 188 (2016).
13. Albayrak, C. *et al. Mol. Cell* **61**, 914–924 (2016).
14. Darmanis, S. *et al. Cell Rep.* **14**, 380–389 (2016).
15. Frei, A.P. *et al. Nat. Methods* **13**, 269–275 (2016).
16. Murphy, K., Travers, P. & Walport, M. *Janeway's Immunobiology* 7th edn. (Garland Publishing, 2008).
17. Robinson, J.P. & Roederer, M. *Science* **350**, 739–740 (2015).
18. Fan, H.C., Fu, G.K. & Fodor, S.P.A. *Science* **347**, 1258367 (2015).
19. Poli, A. *et al. Immunology* **126**, 458–465 (2009).
20. Ferlazzo, G. & Münz, C. *J. Immunol.* **172**, 1333–1339 (2004).
21. Wendt, K. *et al. J. Leukoc. Biol.* **80**, 1529–1541 (2006).
22. Shahi, P., Kim, S.C., Haliburton, J.R., Gartner, Z.J. & Abate, A.R. *Sci. Rep.* **7**, 44447 (2017).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

ONLINE METHODS

See *Protocol Exchange*[23] and **Supplementary Protocol** for a step-by-step protocol for CITE-seq.

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[24]. Antibodies were streptavidin labeled 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 an average of eight binding sites for biotin on each antibody. 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 0.5 M NaCl and 0.02% Tween. Unbound oligo was removed from antibodies using centrifugal filters with a 50 KDa 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.

List of antibodies used for CITE-seq. Antibodies and clones used were CD3e (Clone UCHT1, BioLegend, USA); CD19 (Clone HIB19, BioLegend, USA); CD4 (Clone RPA-T4, BioLegend, USA); CD8a (Clone RPA-T8, BioLegend, USA); CD56 (Clone MEM-188, BioLegend, USA); CD16 (Clone B73.1, BioLegend, USA); CD11c (Clone B-ly6, BD Pharmingen, USA); CCR7 (Clone 150603, R&D Systems, USA); CCR5 (Clone J418F1, BioLegend, USA); CD34 (Clone 581, BioLegend, USA); CD14 (Clone M5E2, BioLegend, USA); CD10 (Clone HI10a, BioLegend, USA); CD45RA (Clone HI100, BioLegend, USA); CD29 (Clone MA1-19105, Thermo Fisher, USA); CD29 (Clone MA5-16707, Thermo Fisher, USA); CD2 (Clone RPA-2.10, BioLegend, USA); CD57 (Clone H-NK1, BioLegend, USA). See **Supplementary Table 2** for a list of antibodies, clones and barcodes used for CITE-seq.

Antibody-oligo sequences. We leverage the DNA-dependent DNA polymerase activity of commonly used reverse transcriptases[25] to convert CITE-seq DNA oligonucleotides into cDNA during reverse transcription at the same time as 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), for further cDNA synthesis. Single cell RNA-seq protocols (including 10x Genomics and Drop-seq) also rely entirely on this activity to append a PCR handle to the 5′ end of full-length cDNAs. The PCR handle 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., 10x Single Cell 3′ v1 uses read 1, while Drop-seq and 10x Single Cell 3′ v2 use read 2). Our proof-of-principle human and 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 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 read 2 handle). BC6: /5AmMC12/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAATNNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

BC12: /5AmMC12/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGTANNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Species mixing, 10x (single cell 3′ version 1, Nextera read1 handle). BC6: /5AmMC12/TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCAATNNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

BC12: /5AmMC12/TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGTANNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

CBMC profiling – (Drop-seq and 10x v2 compatible oligos, containing TruSeq small RNA read 2 handle). v2_BC1: /5AmMC12/CCTTGGCACCCGAGAATTCCAATCACGBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC2: /5AmMC12/CCTTGGCACCCGAGAATTCCACGATGTBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC3: /5AmMC12/CCTTGGCACCGAGAATTCCATTAGGCBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC4: /5AmMC12/CCTTGGCACCCGAGAATTCCATGACCABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC6: /5AmMC12/CCTTGGCACCCGAGAATTCCAGCCAATBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC9: /5AmMC12/CCTTGGCACCCGAGAATTCCAGATCAGBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC10: /5AmMC12/CCTTGGCACCCGAGAATTCCATAGCTTBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC12: /5AmMC12/CCTTGGCACCCGAGAATTCCACTTGTABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC8: /5AmMC12/CCTTGGCACCCGAGAATTCCAACTTGABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC11: /5AmMC12/CCTTGGCACCCGAGAATTCCAGGCTACBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA v2_BC13: /5AmMC12/CCTTGGCACCCGAGAATTCCAAGTCAABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC14: /5AmMC12/CCTTGGCACCCGAGAATTCCAAGTTCCBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

v2_BC5: /5AmMC12/CCTTGGCACCCGAGAATTCCAACAGTGBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

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 min with Fc receptor block (TruStain FcX, BioLegend, USA) to block nonspecific antibody binding. Subsequently, cells were incubated in with mixtures of barcoded antibodies for 30 min 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 min at 4 °C) and supernatant exchange. After the final wash, cells were resuspended at appropriate cell concentration in PBS for Drop-seq[1] or 10x Genomics[3] applications.

CITE-seq on Drop-seq platform. 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 PBMC experiments, cells were loaded at 150 cells/μL. cDNA was amplified for ten 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 for ten 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 desired 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).

CITE-seq on 10x platform. The 10x single-cell run was performed according to the manufacturer's instructions (10x Genomics, USA) with modifications. For the human/mouse mixing experiment (run 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 (run 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 to estimate the true doublet rates (4%) in our experiments and compare these rates to the estimates provided by the equipment manufacturer (~3.1%) (see below). cDNA was amplified for ten 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 for ten 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.

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 (FBS, 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[26]. 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) labeled antibodies and CITE-seq oligo-labeled antibodies from the same monoclonal antibody clone (RPA-T8) targeting CD8a, at concentrations recommended by the manufacturer (1 ug 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, which was 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 were plotted using FlowJo v9 (USA).

Multiparameter flow cytometry. Cells were stained with the following mouse anti-human antibodies, which were purchased from BD Biosciences (USA). Antibodies, clones and fluorophores used were CD3e (clone SK7) Hilyte 750 Allophycocyanin (H7APC), CD4 (clone SK3) Brilliant Blue (BB) 630, CD8a (clone SK1) Phycoerythrin (PE), CD14 (clone M5E2) Brilliant Violet (BV) 750, CD19 (clone HIB19) BV570, CD11c (clone B-ly6) Cyanin5 PE, CD2 (clone RPA-2.10) Brilliant Ultraviolet (BUV) 805, and CD57 (clone, NK-1) BB790. After washing cells in PBS and fixing them 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 were processed with the standard pipeline (Drop-seq tools version 1.12 from McCarroll lab). 10x data from the species mixing experiment were processed using Cell Ranger 1.2 using default parameters, and no further filtering was applied. 10x data from CBMC experiments (v2 chemistry) were processed using the same pipeline as used for our Drop-seq data. Reads were aligned to the human reference sequence GRCh37/hg19 (CD8a FACS comparison) or to an hg19 and mouse reference mm10 concatenation (species mixing experiment, CBMCs).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

doi:10.1038/nmeth.4380

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Drop-seq data of the species mixing experiment were 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 output 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 $\log_{10}$ after adding a pseudocount) that is more than 3 s.d. 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 (minimum counts used: 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% of UMI counts, the assigned species was mouse. Cell barcodes in between 10% and 90% human were considered mixed species. The resulting assignment was 8,005 human, 579 mouse, 33 mixed. 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

$$x_{i,j} = \log\left(\frac{c_{i,j} \times 10{,}000}{m_j}\right)$$

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 556 highly variable genes by fitting a smooth line (LOESS, span = 0.33, degree = 2) to $\log_{10}$(var(UMIs)/mean(UMIs)) as a function of $\log_{10}$(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 are represented as a weighted network with cells being nodes and squared Jaccard similarities as edge weights (based on Euclidian disstance 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[27–29].

CBMC antibody-derived tag 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 log ratio (CLR) transformation[30]. Explicitly, we generated a new CLR-transformed ADT vector y for each cell where

$$y = \mathrm{clr}(x) = \left[\ln\left(\frac{x_1}{g(x)}\right), \ln\left(\frac{x_2}{g(x)}\right), \ldots, \ln\left(\frac{x_5}{g(x)}\right)\right],$$

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, which was perhaps caused by differences in antibody specificity and/or epitope abundance. To compensate for the resulting shifts in the nonspecific 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 s.d. 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%.

Statistical analysis. A **Life Sciences Reporting Summary** is available.

Data availability statement. All raw data generated in this project have been deposited to the Gene Expression Omnibus (GEO) with the accession code GSE100866.

23. Stoeckius, M. & Smibert, P. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.068 (2017).
24. Adler, M., Wacker, R. & Niemeyer, C.M. *Analyst* **133**, 702–718 (2008).
25. Baranauskas, A. *et al. Protein Eng. Des. Sel.* **25**, 657–668 (2012).
26. Breton, G., Lee, J., Liu, K. & Nussenzweig, M.C. *Nat. Protoc.* **10**, 1407–1422 (2015).
27. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. *J. Stat. Mech.* **2008**, P10008 (2008).
28. van der Maaten, L. & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
29. van der Maaten, L. *J. Mach. Learn. Res.* **15**, 1–21 (2014).
30. Aitchison, J. *Math. Geol.* **21,** 787–790 (1989).

doi:10.1038/nmeth.4380