Title
SCITO-seq: single-cell combinatorial indexed cytometry sequencing

Journal
Nature Methods, 18(8)

ISSN
1548-7091

Authors
Hwang, Byungjin
Lee, David S
Tamaki, Whitney
et al.

Publication Date
2021-08-01

DOI
10.1038/s41592-021-01222-3

Peer reviewed

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2022 February 05.

Published in final edited form as:

Nat Methods. 2021 August ; 18(8): 903–911. doi:10.1038/s41592-021-01222-3.

SCITO-seq: single-cell combinatorial indexed cytometry sequencing

Byungjin Hwang[1,2,3,4,5,17], **David S. Lee**[1,2,17], **Whitney Tamaki**[6], **Yang Sun**[1,2], **Anton Ogorodnikov**[1,2], **George C. Hartoularos**[1,2,7], **Aidan Winters**[7], **Bertrand Z. Yeung**[8], **Kristopher L. Nazor**[8], **Yun S. Song**[9,10,15], **Eric D. Chow**[11], **Matthew H. Spitzer**[12,13,14,15], **Chun Jimmie Ye**[1,2,3,4,5,14,15,16,*]

[1]Institute for Human Genetics (IHG), University of California, San Francisco, San Francisco, California, USA

[2]Division of Rheumatology, Department of Medicine, University of California, San Francisco, San Francisco, California, USA

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA

[4]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA

[5]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA

[6]Graduate Program in Pharmaceutical Sciences and Pharmacogenomics, University of California, San Francisco, San Francisco, USA

[7]Graduate Program in Biological and Medical Informatics, University of California, San Francisco, San Francisco, USA

[8]BioLegend Inc, San Diego, CA, USA

[9]Computer Science Division, University of California, Berkeley, Berkeley, CA 94720, USA

[10]Department of Statistics, University of California, Berkeley, CA 94720, USA

[11]Center for Advanced Technology, Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA

[12]Graduate Program in Biomedical Sciences, University of California, San Francisco, San Francisco, CA 94143, USA

[*]Correspondence should be addressed to C.J.Y (jimmie.ye@ucsf.edu).

[17]these authors contributed equally to this work

Author contributions

C.J.Y., B.H and D.S.L conceived the experiments, B.H. and D.S.L. designed and conducted the experiment(s), B.Y. and K.L.N. kindly provided antibodies for commercial compatibility experiments. C.J.Y., B.H., D.S.L., W.T., A.O., G.H., A.W., Y.S.S., Y.S., E.D.C., and M.H.S. analyzed the results. All authors reviewed the manuscript.

Code availability

All code used to perform simulations and generate figures can be found at the following websites (https://github.com/yelabucsf/SCITO-seq, https://doi.org/10.5281/zenodo.4988182). Useful project related cost calculation website can be found also at (https://yelabtools.herokuapp.com/scSeqCostCalc/scito.html).

[13]Departments of Otolaryngology and Microbiology and Immunology, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA 94143, USA

[14]Parker Institute for Cancer Immunotherapy, San Francisco, CA 94158, USA

[15]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

[16]J. David Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA

Abstract

The development of DNA-barcoded antibodies to tag cell-surface molecules has enabled the use of droplet-based single cell sequencing (dsc-seq) to profile protein abundances from thousands of cells simultaneously. Compared to flow and mass cytometry, high per cell cost of current dsc-seq-based workflows precludes their use in clinical applications and large-scale pooled screens. Here, we introduce SCITO-seq, a workflow that uses splint oligonucleotides (oligos) to enable combinatorially indexed dsc-seq of DNA-barcoded antibodies from greater than 10^5 cells per reaction using commercial microfluidics. By encoding sample barcodes into splint oligos, we demonstrate that multiplexed SCITO-seq produces reproducible estimates of cellular composition and surface protein expression comparable to mass cytometry. We further demonstrate two modified splint oligo designs that extend SCITO-seq to achieve compatibility with commercial DNA-barcoded antibodies and simultaneous expression profiling of the transcriptome and surface proteins from the same cell. These results demonstrate SCITO-seq as a flexible and ultra high-throughput platform for sequencing-based single-cell protein and multimodal profiling.

Introduction

The use of DNA to barcode physical compartments and tag intracellular and cell-surface molecules has enabled the use of sequencing to efficiently profile the molecular characteristics of thousands of cells simultaneously. While initially applied to measuring the abundances of mRNAs[1,2] and identifying regions of accessible chromatin[3], recent developments in DNA-tagged antibodies have created new opportunities to use sequencing to measure the abundances of surface[4,5] and intracellular proteins[6].

Sequencing-based protein profiling of single cells has several advantages over flow and mass cytometry. First, the number of surface proteins that can be measured by DNA-tagged antibodies is exponential in the number of bases in the tag. In theory, the entire cell surfaceome can be targeted and is only limited by the availability of antibodies. In practice, panels targeting hundreds of proteins are already commercially available[4,7]. This contrasts with other cytometry methods where the number of proteins targeted is limited by the overlap in the emission spectrums of fluorophores (flow: 4–48) or the number of unique masses of metal isotopes that can be chelated by commercial polymers (CyTOF: ~50)[8,9]. Second, sequencing-based protein profiling can readily read out all antibody-derived tags (ADTs) with one sequencing experiment instead of multiple rounds of signal separation and detection, significantly reducing the time and sample input for profiling large panels. Third, additional molecules can be profiled from the same cell enabling multimodal profiling

of cells[4,10,11]. Finally, sequencing is amenable to encoding orthogonal information using additional DNA barcodes creating opportunities for clinical applications and pooled screens that barcode cells using natural variation[12], synthetic sequences[13,14], or sgRNAs[15,16].

Despite the opportunities, the major limitation in sequencing-based single-cell protein profiling[4,7] is the high cost associated with profiling each cell, thus precluding its use for phenotyping population cohorts or in large-scale screens where millions of cells per experiment would be desirable. Like other single-cell sequencing assays, the total cost per cell for protein profiling is divided between costs associated with library construction and library sequencing. Because the number of protein molecules per cell is 2–6 orders of magnitude higher than mRNA molecules for the same gene[17] and the use of targeting antibodies limits the number of proteins profiled, sequencing ADTs generally requires fewer number of reads and thus has a lower sequencing cost per cell compared to sequencing the transcriptome. However, library construction costs for commercial microfluidics-based single-cell sequencing platforms[18] are high regardless of the modality. Furthermore, conjugation of modified DNA sequences to antibodies can incur a high one time cost[4,5]. Thus, for sequencing-based protein profiling to be a compelling strategy for ultra high-throughput cellular phenotyping, there is a need to develop workflows that minimize library preparation costs without incurring significant antibody conjugation costs.

Single-cell combinatorially indexed (SCI) sequencing is an alternative approach to commercial microfluidics for massively parallel single-cell sequencing by labeling sequential rounds of physical compartmentalization with DNA barcodes. While initially applied to mRNA and chromatin accessibility sequencing, a recent study[19] demonstrated surface protein profiling using SCI-sequencing. However, standard SCI-sequencing requires more than two rounds of combinatorial indexing to profile 10^5 – 10^6 cells[19–23]. Coordinating multiple rounds of chemical and enzymatic reactions to ensure DNA-barcoded antibodies remain bound to surface proteins during each round is experimentally challenging. Furthermore, profiling both mRNA and protein using SCI-sequencing would require compatibility of multiple cycling conditions which may be highly inefficient. Indeed, the published approach[19] only measured 14 mRNA transcripts and 1 protein simultaneously for 5×10^4 cells with 4 rounds of indexing and 60 wells per round.

Here, we introduce single-cell combinatorial indexed cytometry sequencing (SCITO-seq), a simple and cost-effective procedure for single-cell protein profiling that combines split-pool indexing and droplet-based single cell sequencing. Our approach is based on the key insight that the large number of droplets produced by commercial microfluidic devices (e.g. up to 10^5 for 10X Genomics[18]) can be used as a second round of physical compartments for two-step combinatorial indexing. While two-step SCI workflows have been recently described for ATAC-seq[24] and RNA-seq[25], they have not been implemented for protein and multimodal profiling to the best of our knowledge. For SCITO-seq, we introduce a strategy using universal conjugation of antibodies followed by pooled hybridization of splint oligos to minimize the cost of generating multiple pools of DNA-tagged antibodies. Tagged antibodies are then used to stain cells in respective pools prior to high-concentration loading and library construction utilizing commercial microfluidics. After sequencing, protein expression profiles for cells simultaneously encapsulated in a single droplet can be resolved

with pool barcodes. In mixed species and mixed individual experiments, we demonstrate that SCITO-seq achieves a throughput on the order of 10^5 cells per microfluidic reaction with low collision rates tunable by the number of antibody pools. Then, we applied SCITO-seq to profile peripheral blood mononuclear cells (PBMCs) using panels of 28 and 60 antibodies generating reproducible estimates of cellular composition and surface protein expression comparable to mass cytometry (CyTOF). We further demonstrate the flexibility of SCITO-seq by modifying the splint oligo design to achieve compatibility with commercial DNA-barcoded antibody panels and an existing workflow[25] to enable combinatorially indexed multimodal profiling of the whole transcriptome and surface proteins.

Results

Theory and design of SCITO-seq

Because cell loading in microfluidic devices is governed by a Poisson process, all dsc-seq workflows recommend low loading concentrations to maintain a low collision rate, defined as the frequency of droplets containing multiple cells. For the 10X Genomics single-cell sequencing platform, Poisson loading at the recommended $2 \times 10^3 – 2 \times 10^4$ cells per channel result in cell recovery rates of 50–60%[18,25] and collision rates of 1–10% (Extended Data Fig. 1a). At these recommended loading concentrations, 97–82% of droplets do not contain a cell, resulting in suboptimal usage of reagents (Extended Data Figures 1b and c). One approach to decrease the library preparation cost of dsc-seq is to “barcode” samples using either natural genetic variants[12,26] or synthetic DNA molecules[13,14,27] prior to pooled loading at higher numbers of cells per channel (e.g. loading $5 \times 10^4 – 8 \times 10^4$ cells), thus reducing the proportion of droplets without a cell (e.g. 65–45%). Because simultaneous encapsulation of some cells within a droplet can be detected by the co-occurrence of different sample barcodes (e.g., genetic profiles or synthetic DNA tags), sample multiplexing increases the number of cells recovered per channel while maintaining a low effective collision rate that is tunable by the number of sample barcodes. However, as collision events can only be detected but not resolved into usable single-cell data, the maximum loading concentration that minimizes total cost is ultimately limited by the overhead incurred for sequencing collided droplets.

Here we propose a simple two round single-cell combinatorially indexed (SCI) experimental workflow, SCITO-seq, which combinatorially indexes single cells using DNA-tagged antibodies and microfluidic droplets to enable cost-effective profiling of cell-surface proteins that is scalable to $10^5 – 10^6$ cells per microfluidic channel (Fig. 1a). First, each antibody is conjugated with an antibody-specific amine-modified oligo that also serves as a hybridization sequence (Ab Handle, 20bp). Second, titrated antibodies are mixed and aliquoted into pools before the addition of splint oligos, each containing a compound barcode consisting of an antibody barcode and a pool barcode (Ab+PBC). Each splint oligo also contains complementary sequences to the Ab Handle and to the gelbead-bound sequences (Fig. 1b). The design of the antibody and gelbead hybridization sequences can be customized for compatibility with commercial antibody conjugation chemistries and droplet-based chemistries, respectively. Third, cells are aliquoted into pools and stained with antibodies hybridized with pool-specific barcodes. Fourth, the stained cells are pooled and

loaded using a commercial dsc-seq platform at loading concentrations tunable to the targeted collision rate followed by library construction incorporating unique molecular identifiers (UMI) and droplet barcodes (DBC). Finally, after sequencing the antibody derived tags (ADTs), the surface protein expression profiles of simultaneously encapsulated cells within a droplet (multiplets) can be resolved using the combinatorial index of Ab+PBC and DBC.

Key to the scalability of SCITO-seq is the recognition that Poisson loading naturally limits the number of cells within a droplet even at very high loading concentrations. Thus, indexing cells using a small number of antibody pools will be sufficient to ensure that the combinatorial index (Ab+PBC and DBC) will resolve cells at low collision rates. Theoretically, given P pools, C cells loaded, D droplets formed, the collision rate is given as $\mathbb{P}[\text{Collision}] = 1 - e^{-\frac{C}{D}}\left[1 + \frac{C}{PD}\right]^{P}$ while the rate of empty droplets is given by $\mathbb{P}[\text{Empty}] = e^{-\frac{C}{D}}$ (see Methods for derivation). Our derivation of the collision rate differs from previous estimates derived from the classical birthday problem[25], which did not account for encapsulation of more than two cells within the same droplet. These closed form derivations of the collision and empty droplet rates are nearly identical to those obtained from simulations (Extended Data Fig. 1d). For example, assuming 6×10^4 droplets are formed and 1.82×10^5 cells are loaded, 84% of droplets contain at least one cell but only 4.4% of droplets contain greater than four cells. To yield 10^5 resolved cells at a collision rate of 5%, only 10 antibody pools would be needed. Given current costs, this corresponds to a total cost of 3.1 ¢/cell. Note that as the library preparation cost quickly diminishes for SCITO-seq with increasing number of pools, the total cost per cell is dominated by antibody costs. Therefore, while 384 pools achieve the maximal 12-fold reduction in cost compared to standard sequencing-based single-cell protein profiling (2.2 vs 26 ¢/cell), 10 antibody pools can already achieve an 8-fold reduction in cost (3.1 vs 26 ¢/cell) with minimal additional experimental complexity (Extended Data Fig. 1e).

Dynamic range and validation of SCITO-seq

We first assessed the sensitivity and dynamic range of SCITO-seq by profiling two pools of PBMCs from a single donor stained with equal proportions of a fluorescently labeled CD8A antibody and one of two pool-barcoded SCITO-seq CD8A antibodies. After washing unbound antibodies, the stained pools were equally mixed and FACS sorted into four bins based on CD8A expression. Each sorted bin was reanalyzed by flow cytometry or tagged with a unique hashtag. The hashtagged cells were pooled and 30,000 cells were loaded and processed using the 10X Chromium 3' V3 kit for ADT library construction and sequenced to a depth of 3,647 reads/cell (Fig. 1d, See Methods). The distribution of CD8A ADT counts was highly similar between the two SCITO-seq pools (Pool 1: 2.8+/−1.2, 4.6+/−0.4, 6.5+/−0.6, 7.3+/−0.5; Pool2: 2.3+/−1.1, 4.3+/−0.5, 6.4+/−0.5, 7.2+/−0.2). Compared to flow cytometry, SCITO-seq produced similar coefficients of variation across the four bins (SCITO-seq: 0.42, 0.11, 0.08, 0.05; flow: 0.30, 0.15, 0.07, 0.04) at a slightly lower dynamic range (SCITO-seq: 0–9660 UMIs, flow: 0–22610 intensity units) likely due to limited sequencing. These results suggest SCITO-seq has similar noise profiles to flow cytometry across a dynamic range of four orders of magnitude with little pool-to-pool variation.

To assess the feasibility of SCITO-seq, we performed a mixed species experiment with human (HeLa) and mouse (4T1) cell lines in five pools, each stained with anti-human CD29 (hCD29) and anti-mouse CD29 (mCD29) antibodies hybridized with pool-specific barcodes (Fig. 1e). After mixing the stained pools at equal proportions, 10^5 cells were loaded and processed using the 10X Chromium 3' V3 kit for ADT and mRNA library construction. The resulting libraries were sequenced to recover 38,504 cell-containing droplets (CCDs) at a depth of 2,909 reads per CCD for ADT and 25,844 reads per CCD for mRNA. We merged and analyzed ADT counts for each antibody across pools to mimic standard sequencing-based protein profiling (e.g., CITE-seq[4]). Of the 38,504 CCDs, 40.6% contained only mCD29 ADTs, 35.7% contained only hCD29 ADTs, and 21.9% contained ADTs from both species, which we labeled as between-species multiplets (Fig. 1f). These estimates were consistent with results from analyzing the transcriptomic data: 42.7% CCDs contained mouse transcripts, 33.9% contained human transcripts, and 23.3% contained transcripts from both species (Extended Data Fig. 2a). ADT counts from both between- and within-species multiplets could be resolved by utilizing the DBC and Ab+PBCs, reducing the collision rate from an estimated 51% to 8.8% (expected 6.3%) (Fig. 1f, Extended Data Fig. 2b) without significant pool-to-pool variation (Extended Data Fig. 2c). After resolving multiplets, protein expression profiles were obtained for 46,295 cells, a 3.7-fold throughput increase over the expected ~12,500 cells from standard workflows at the same collision rate (Fig. 1g). Further, a two-pool SCITO-seq experiment produced similar results to direct conjugation of four different Ab+PBC barcodes suggesting that both within and between pool splint oligo contamination rates are low (Extended Data Fig. 2d and e).

We next tested the cellular throughput of SCITO-seq and its applicability to resolve quantitative differences in cellular composition based on surface protein expression. We isolated and mixed primary $CD4^+$ T and $CD20^+$ B cells from two donors at a ratio of 5:1 (T:B) for donor 1 and 1:3 (T:B) for donor 2. The mixed cells were aliquoted into five pools and each stained with pool-barcoded anti-CD4 and anti-CD20 antibodies (Fig. 1h). Stained pools were mixed at equal ratios, 2×10^5 cells were loaded and processed using the 10X Chromium 3' V3 kit, and the resulting ADT and mRNA libraries sequenced to recover 58,769 post-processing CCDs. Merging the ADT data across the five pools, anti-CD4 and anti-CD20 antibodies stained the expected cell populations as defined by the transcriptomic data. Based on the ADTs, 40% of CCDs were estimated to be between cell-type multiplets, which was consistent with estimates from the transcriptomic analysis (49.6%; Fig. 1i, Extended Data Fig. 3a). Based on genetic demultiplexing that leverages genetic variants captured in the transcriptomic data, an additional 30% of CCDs were estimated to be within cell-type multiplets for a total multiplet rate of 70%. After resolving both between and within cell-type multiplets using the combinatorial index of Ab+PBC and DBC (Fig. 1j, Extended Data Fig. 3b), a total of 116,827 cells were obtained at a collision rate of 25%, effectively increasing the throughput by 4.0-fold over standard workflows at the same collision rate. Note that both the multiplet rates ($R = 0.97$, $P < 0.01$) and the co-occurrence rates of SCITO-seq antibodies from different pools ($R = 0.93$, $P < 0.01$) were highly correlated between the expected and observed values (Extended Data Fig. 3c) suggesting that the encapsulation of multiple cells within a CCD is not biased for specific pools or cell types. We next assessed if SCITO-seq can capture unequal distributions of B

and T cells from the two donors, especially from CCDs that encapsulated multiple cells. For this analysis, we focused only on the 45,240 CCDs predicted to contain cells from only one donor based on genetic demultiplexing (donor 1: 25,630, donor 2: 19,610). Within CCDs with only one SCITO-seq pool barcode detected (singlets), the estimated proportions of T and B cells (T: B_{200K}: 5.0:1 for donor 1 and 1:2.8 for donor 2) were consistent with the expected proportions for each of the two donors and estimates obtained from the transcriptomic data (Extended Data Fig. 3d). Approximately the same proportions were estimated in CCDs with multiple pool barcodes (multiplets) (T: B_{200K} 4.0:1 for donor 1 and 1:2.9 for donor 2).

Ultra high-throughput phenotyping of PBMCs with SCITO-seq

To demonstrate the applicability of SCITO-seq for ultra high-throughput multidimensional cellular phenotyping using surface proteins, we profiled peripheral blood mononuclear cells (PBMCs) from two healthy donors using a panel of 28 monoclonal antibodies across 10 pools. After staining, pooling, and processing 2×10^5 cells in a single channel using the 10X Chromium 3' V3 kit, the resulting ADT and mRNA libraries were sequenced and analyzed to obtain 49,510 post-filtering CCDs (Fig. 2a). Each of the 10 SCITO-seq pool barcodes was detected in a subset of CCDs at levels significantly higher than other pool barcodes suggesting a high signal-to-noise ratio to resolve multiplets (Extended Data Fig. 4a). A total of 93,127 cells were resolved at a collision rate of 8.5%, increasing the throughput by 10-fold over standard workflows at the same collision rate consistent with simulations (Extended Data Fig. 1d).

We separately analyzed the mRNA, merged ADT, and resolved ADT data by normalizing the raw counts, performing dimensionality reduction, constructing a k-nearest neighbor graph, and clustering using the Leiden algorithm (see Methods). Leiden clusters based on either merged ADT or RNA counts (Fig. 2a, Extended Data Fig. 4b and 5a) were poorly differentiated in Uniform Manifold Approximation and Projection (UMAP) space due to the high loading concentrations as 69% of CCDs are expected to contain more than one cell. Strikingly, Leiden clusters based on resolved ADT data resulted in 17 distinct clusters in UMAP space which could each be annotated based on the expression of lineage-specific surface markers including: eight clusters of the myeloid lineage, naïve and memory $CD4^+$ and $CD8^+$ T cells, natural killer (NK) cells, B cells and gamma delta T cells (gdT) (Fig. 2b, Extended Data Fig. 5b). Notably, naive ($CD45RA^+$) and memory ($CD45RO^+$) $CD4^+$ and $CD8^+$ T cells emerged as separate clusters which can be difficult to distinguish based on transcriptomic data due to low mRNA abundances of lineage markers (e.g. CD4) and inability to infer isoforms (e.g. CD45RO) directly from 3' sequencing data[18]. Indeed, analyzing the transcriptomes of CCDs likely containing only a single cell showed limited separation of naive and memory cells (Extended Data Fig. 5c and d).

We next assessed the accuracy of SCITO-seq for quantitative immune phenotyping by comparing the compositional estimates obtained from CCDs containing a single SCITO-seq pool barcode (singlets) versus those containing multiple pool barcodes (multiplets). We focused the analysis on CCDs containing cells from one donor predicted based on genetic demultiplexing. UMAP projections for resolved cells originating from singlets or multiplets

were qualitatively similar (Fig. 2c), suggesting that higher rates of encapsulation do not generate technical artifacts. The frequency estimates of the 16 immune populations detected from singlets and multiplets were more similar from the same donor (average cosine similarity: 0.98 [donor 1], 0.97 [donor 2]; Fig. 2d) than between different donors (average cosine similarity: 0.83). We further compared the SCITO-seq data to mass cytometry (CyTOF) data generated using the same antibodies conjugated to metal isotopes (Extended Data Fig. 6a). Joint clustering of the CyTOF and SCITO-seq data produced qualitatively similar UMAP projections (Fig. 2c) and the frequency estimates of jointly annotated cell types were highly similar between assays for each donor (average cosine similarity: 0.95 [donor 1], 0.93 [donor 2]) (Fig. 2d and Extended Data Fig. 6b).

Expanding SCITO-seq to custom and commercial antibody panels

To further demonstrate the flexibility and scalability of SCITO-seq beyond the number of markers detectable by flow and mass cytometry[9,28], we evaluated the performance of SCITO-seq using a 60-plex custom antibody panel and a commercial Totalseq-C (TSC) 165-plex antibody panel (Extended Data Fig. 7a). To achieve compatibility with the commercial TSC panel which utilizes 5' conjugation, we redesigned splint oligos that would hybridize to each antibody-specific 15bp barcode. For each panel, we stained 10 donors in separate pools utilizing the pool barcode as a sample label, mixed the pools, and processed 4×10^5 cells using the 10X Chromium 3' V3 kit targeting the recovery of 2×10^5 cells per experiment. In the 60-plex experiment, we recovered 69,733 CCDs and resolved 219,063 cells (Fig. 3a, 3b, Extended Data Fig. 7b, and c) at a collision rate of 18.7%. In the TSC experiment, we recovered 66,774 CCDs and resolved 203,838 cells (Fig. 3c and 3d) at a collision rate of 14.1%. Further, the counts of isotype control antibodies were consistently low (Extended Data Fig. 7d) suggesting non-specific background staining is not an appreciable factor (Extended Data Fig. 7e and f). Note that even when loading 4×10^5 cells, 20-fold higher than recommended, we did not observe a decrease in the number of UMIs recovered per cell versus the number of cells per CCD suggesting that reagents are not yet a limiting factor at high loading concentrations (Fig. 3e). Further, the observed multiplet rates were highly correlated (60-plex; R=0.99, P-value < 0.001, TSC; R=0.92, P-value < 0.001) with those obtained from simulation (Fig. 3f).

After additional removal of collided barcodes that could be detected based on co-occurrence of lineage markers (See Methods), 175,930 and 175,000 cells were obtained in the 60-plex and 165-plex experiments respectively. After count normalization, dimension reduction, and k-nearest neighbor graph construction, the cells were clustered using the Leiden algorithm into 26 and 19 clusters respectively and visualized in UMAP space (Fig. 3a, c). The expected lymphoid and myeloid cell types were annotated with lineage markers (Fig. 3b, d). Compared to the 28-plex dataset, higher dimensional phenotyping enabled the identification of low frequency cell types such as two populations of conventional dendritic cells (cDC1s and cDC2s) distinguished by the expression of CD141, CD370, CD1C and plasmacytoid dendritic cells (pDCs) by the expression of CD123, CD303 and CD304[29] (Fig. 3b, d and Extended Data Fig. 8a).

SCITO-seq creates new opportunities for multiplexed profiling as pool barcodes can be used to directly label samples obviating the need for orthogonal barcoding (Extended Data Fig. 8b). Encouragingly, similar to previous observations (Extended Data Fig. 3b and 5b), a pairwise analysis across all antibodies demonstrated no significant correlation across pools (Extended Data Fig. 8c and d) suggesting minimal pool-specific effects that may confound quantitative comparisons of multiple samples. We assessed the performance of multiplexed SCITO-seq by comparing the compositional and surface expression variation between the 60-plex and 165-plex experiments since the same 10 individuals were processed. The frequency estimates of T, NK, B, and myeloid cell populations were highly correlated between experiments (R=0.98–0.99, P-value < 0.001; Fig. 3g). Further, the expression levels of 43 surface proteins profiled in both experiments were highly correlated when all PBMCs (Fig. 3h) and when only individual cell types (Fig. 3i) were analyzed. These results suggest that SCITO-seq can capture both interindividual variability in composition estimated from the entire marker panel and expression of individual surface proteins that do not generally emerge as distinct UMAP clusters.

Combinatorial-indexed transcriptomic and proteomic profiling

Finally, we sought to enable ultra high-throughput multimodal profiling of the transcriptome and surface proteins by integrating SCITO-seq with the recently published scifi-RNA-seq workflow[25]. Scifi-RNA-seq combinatorially indexes mRNAs by *in-situ* reverse transcription utilizing pool-specific RT-primers followed by ligation of pool-barcoded cDNA molecules with a DBC utilizing the 10X single-cell ATAC-seq kit (scATAC-seq). To first enable compatibility of SCITO-seq with the scATAC-seq chemistry, we modified the gelbead hybridization sequence of the splint oligo to be complementary to the scATAC-seq gelbead sequence. As a proof of principle, we applied SCITO-seq utilizing the modified splint oligo and the scATAC-seq kit to profile PBMCs from one donor in five pools with 12 broad phenotyping surface markers (See Methods). A total of 5×10^4 cells were loaded and sequenced to recover 21,460 cells. Analysis of the data identified the expected clusters of T, B, myeloid, and NK cells expressing the canonical surface proteins demonstrating the compatibility of SCITO-seq with scATAC-seq chemistry (Extended Data Fig. 9a).

Scifi-RNA-seq utilizes a bridge oligo to facilitate the ligation of DBCs and requires a number of cycling conditions that is not directly compatible with SCITO-seq. To enable multimodal profiling, we modified the orientation of the SCITO-seq splint oligo and designed an orthogonal bridge oligo to assist capture and ligation of SCITO-seq ADTs to the 10X scATAC-seq gelbead capture sequence (Fig. 4a and Extended Data Fig. 9b). This modification allowed for the addition of a DBC to ADTs while minimizing the competition between bridge oligo capture of cDNA and ADT molecules utilizing the existing scifi-RNA-seq protocol. As a proof of principle, we stained a mixture of four human cell lines (LCL, NK-92, HeLa, Jurkat) and one mouse cell line (4T1) with antibodies hybridized with the modified SCITO-seq oligos targeting six surface proteins in five pools (Fig. 4a). 3×10^4 stained cells were processed using the scifi-RNA-seq workflow including both ADT and mRNA bridge oligos resulting in 10,439 cells sequenced based on the ADT data. After pre-processing, we obtained an average of 310 UMIs per cell for the mRNA library (average 146 genes/cell) and an average of 550 UMIs per cell for the ADT library. Each mRNA and

ADT pool barcode was detected in a subset of CCDs that minimally contained barcodes from other pools suggesting high signal to noise ratio in resolving both modalities (Fig. 4b and c). This is confirmed by the near equal distribution of resolved cells expressing either human or mouse CD29 ADTs (Gini index of 0.12) (Fig. 4d). After normalization of the ADT counts, dimensionality reduction, and k-nearest neighbor graph construction, 5 clusters were identified using Leiden clustering and visualized in UMAP space (Fig. 4e). Cells with the highest average expression of cell line-specific transcript markers generally corresponded to the clusters defined using surface protein markers (Fig. 4f). NK-92 specific transcripts were notably less prominently expressed in the the NK-92 cluster defined by ADT likely due to the lower mRNA capture efficiency within these cells (168 UMIs per cell). To further assess the consistency of transcriptomic and ADT data, we constructed a UMAP using the mRNA data and overlaid the ADT clusters annotations. This analysis produced qualitative enrichment (Extended Data Fig. 9c and d) which was quantitatively confirmed including the overlap between NK-92 cells defined using the mRNA and ADT data (Fig. 4g, See Methods). These results demonstrate a provisional integration of SCITO-seq with scifi-RNA-seq that has the potential for ultra high-throughput multimodal profiling of RNA and proteins from the same cells.

Discussion

Current cost for sequencing-based protein profiling using commercial microfluidics is three to four orders of magnitude higher than conventional cytometry methods. While antibody conjugation costs remain similar across platforms, this high cost can be attributed to costs associated with sequencing library construction due in part to the low cell throughput using commercial microfluidic instruments. To reduce the cost of library construction and to maximize cell yields, we developed SCITO-seq, a simple two-step combinatorial indexing approach that enables scalable profiling of > 150 surface proteins from > 10^5 cells in a single microfluidic reaction.

While an approach using standard SCI has been reported for single-cell sequencing-based protein profiling[19], combinatorial indexing using multiple 96/384 well plates is both experimentally laborious and susceptible to noise introduced in each round from free oligos non-specifically binding to cells. By leveraging commercial microfluidics for combinatorial indexing, SCITO-seq presents two distinct advantages due to the natural limited dilution using microfluidic cell loading. First, a two-round split-pool indexing workflow is amenable to the removal of unbound oligos after splint oligo hybridization to minimize background noise. Second, the large number of droplets produced allows for more cells to be sequenced using fewer initial pooling reactions. Indeed, SCITO-seq using ten pool barcodes can profile 10^5 cells, while a conventional SCI method would require four rounds of 60 wells of ligation[19] to profile 5×10^4 cells. To further increase throughput while maintaining its ease of use, SCITO-seq can be combined with sample hashtags[13,14] to enable detection and removal of multiple encapsulated cells with the same pool barcode.

In addition to throughput, we demonstrate the flexibility and scalability of the splint oligo design of SCITO-seq in several applications. First, by leveraging pool-specific splint oligos for sample multiplexing, we demonstrate the feasibility of SCITO-seq to be used

for population-scale phenotyping by minimizing per sample costs and experimental batch effects. Second, we show the scalability of SCITO-seq by profiling 175,000 cells using a commercial panel of 165 antibodies in a single microfluidic channel. This enables extensive phenotyping of cells and identification of low frequency cell types. Third, the splint oligo can be modified to be compatible with a number of chemistries to enable extensions of SCITO-seq for multimodal profiling.

As a proof of principle, we designed a splint oligo compatible with the 10X scATAC-seq chemistry which is also used for scifi-RNA-seq, a novel droplet-based SCI method for profiling mRNA. We further designed an orthogonal bridge oligo for ligation to maximize the efficiency of capture for both mRNA and ADT molecules, enabling simultaneous profiling of the whole transcriptome and surface proteins. As new highly efficient and easy to implement SCI methods begin to emerge[24,25], we envision that SCITO-seq could be further integrated with these methods for new multimodal profiling opportunities. However, as the number of cells profiled increases, a major limitation of any co-assay will be sequencing costs associated primarily with profiling the transcriptome or epigenome. The scalability delivered by SCITO-seq allows for emerging large-scale analyses of millions of cells such as pooled perturbation screens with barcoded perturbations, single-cell cytometry across large population and disease cohorts, and immune repertoire and phenotyping studies of clonal evolution.

Methods

Closed from derivation of collision and empty droplet rates

Suppose there are P pools of cells. For pool p, cells arrive according to a Poisson point process with rate $\lambda_p > 0$ (abbreviated PPP(λ_p)), where the unit of time corresponds to the inter-arrival time of droplets. In the most general formulation, we assume that the point processes for different pools are independent. Further, we assume that a gelbead (respectively, a cell) has probability ρ_p^b (respectively, ρ_p^c) of successfully being encapsulated into a droplet. Therefore, by Poisson thinning, the arrival of cells follows $\text{PPP}(\rho_p^c \lambda_p)$.

We are interested in the probability of the event (called collision) that a droplet contains two or more cells from the same pool. Let N_p denote the number of cells from pool p successfully loaded into a droplet. Then, $N_1, N_2, \ldots, N_p$ where $N_p \sim \text{Poisson}(\rho_p^c \lambda_p)$, are independent random variables, and $\mathbb{P}[\text{Droplet Collision}]$ can be computed as $1 - \mathbb{P}[\text{No Droplet Collision}]$. Here $\mathbb{P}[\text{No Droplet Collision}]$ represents a probability that every droplet contains ≤ 1 pool barcode. Therefore, we derive:

$$\begin{aligned}
\mathbb{P}[\text{Droplet Collision}] &= 1 - \mathbb{P}[\text{No Droplet Collision}] \\
&= 1 - \mathbb{P}\left[(N_1 \leq 1) \cap (N_2 \leq 1) \cap \cdots \cap (N_p \leq 1)\right] \\
&= 1 - \mathbb{P}\left[(N_1 \leq 1)\mathbb{P}(N_2 \leq 1) \cdots \mathbb{P}(N_p \leq 1)\right] \\
&= 1 - \prod_{p=1}^{P} \left[e^{-\rho_p^c \lambda_p}(1 + \rho_p^c \lambda_p)\right]
\end{aligned}$$

where the third equality follows from independence.

Next we wish to compute the conditional probability of “Droplet Collision” given “Non-empty Droplet”. First, note that the probability that a droplet contains a cell at a given observation is $\mathbb{P}[\text{Non-empty Droplet}] = 1 - \mathbb{P}[\text{Empty Droplet}]$, where

$$\begin{aligned}\mathbb{P}[\text{Empty Droplet}] &= \mathbb{P}\left[(N_1 = 0) \cap (N_2 = 0) \cap \cdots \cap \left(N_p = 0\right)\right] \\ &= \prod_{p=1}^{P} e^{-\rho_p^c \lambda_p}\end{aligned}$$

If there are D droplets formed and a total of C cells loaded evenly across the P pools (i.e., there are $\frac{C}{P}$ cells per pool), then $\lambda_p = \frac{C}{PD\rho_p^b}$ for all pools $p = 1, 2, \ldots, P$. If we further assume that $\rho_p^c = \rho^c = 1$ for all $p = 1, 2, \ldots, P$, then $\mathbb{P}[\text{Droplet Collision}]$ and $\mathbb{P}[\text{Empty Droplet}]$ simplify as:

$$\begin{aligned}\mathbb{P}[\text{Droplet Collision}] &= 1 - e^{-\frac{C}{D}}[1 + \frac{C}{PD}]^P \\ \mathbb{P}[\text{Empty Droplet}] &= e^{-\frac{C}{D}}\end{aligned}$$

Finally, the conditional probability of barcode collision in a droplet given that the droplet is non-empty can be found as

$$\begin{aligned}\mathbb{P}[\text{Droplet Collision} \mid \text{Non-empty Droplet}] &= \frac{\mathbb{P}[\text{Droplet Collision}]}{1 - \mathbb{P}[\text{Empty Droplet}]} \\ &= \frac{1 - e^{-\frac{C}{D}}\left[1 + \frac{C}{PD}\right]^P}{1 - e^{-\frac{C}{D}}}.\end{aligned}$$

A second collision rate we can calculate is the cell barcoding (droplet barcode + pool barcode) collision rate which can be computed as the conditional probability that a particular pool $p \in \{1, 2, \ldots, P\}$ has a collision in a given droplet, given that the droplet contains at least one cell from that pool. If we assume that there are D droplets formed and a total of C cells are distributed evenly across P pools, then we obtain:

$$\mathbb{P}[\text{Collision in pool } p \mid \text{Droplet contains at least one cell from pool } p] = \frac{1 - e^{-\frac{C}{PD}}\left[1 + \frac{C}{PD}\right]}{1 - e^{-\frac{C}{PD}}},$$

for all $p \in \{1, 2, \ldots, P\}$.

The above conditional probability is related to the proportion of pools with a collision in a given droplet, relative to the total number of pools each with at least one cell represented in the droplet. More precisely,

$$\frac{\mathbb{E}[\text{Number of pools with a collision in a droplet}]}{\mathbb{E}[\text{Number of pools represented at least once in a droplet}]} = \frac{P \cdot \left[1 - e^{-\frac{C}{PD}}\left(1 + \frac{C}{PD}\right)\right]}{P \cdot \left[1 - e^{-\frac{C}{PD}}\right]} = \frac{1 - e^{-\frac{C}{PD}}\left[1 + \frac{C}{PD}\right]}{1 - e^{-\frac{C}{PD}}}.$$

Simulations of collision and empty droplet rates

For simulating the collision rates and empty droplet rates, we assumed a cell recovery rate of 60% and 10^5 droplets are formed per microfluidic reaction resulting in $D = 6 \times 10^4$. For C cells loaded, cell containing droplets are simulated using a Poisson process where $\lambda = \frac{C}{D}$. Assuming each simulated droplet i contains γ_i cells, we then compute the number of pool barcodes not tagging a cell in each droplet as:

$$BC0_i = P\left(1 - \frac{1}{p}\right)^{\gamma i}$$

the number of pool barcodes tagging exactly one cell as:

$$BC1_i = C_i\left(1 - \frac{1}{p}\right)^{\gamma i - 1}$$

and the number of pool barcodes tagging greater than one cell as:

$$BCN_i = P - BC0_i - BC1_i$$

The conditional collision rate is estimated as:

$$\widehat{\mathbb{P}}[\text{Collision in pool } p \,|\, \text{Droplet contains at least one cell from pool } p] = \frac{\sum_i^c BCN_i}{\sum_i^c BN_i + \sum_i^c BC1_i}$$

Primary antibody oligonucleotide conjugation

For the species mixing experiment, anti-human CD29 and anti-mouse CD29 antibodies were purchased from Biolegend (cat. 303021, 102235) and conjugated per antibody using a ThunderLink kit (Expedeon cat. 425–0000) to distinct 20 bp 3' amine-modified HPLC-purified oligonucleotides (IDT) to serve as hybridization handles. Antibodies were conjugated at a ratio of 1 antibody to 3 oligonucleotides (oligos). In parallel, oligos similar to current antibody sequencing tags were directly conjugated at the same ratio for comparison. Sequences for the hybridization oligonucleotides and directly conjugated oligos were designed to be compatible with the 10x feature barcoding system by introducing a reverse complementary sequence to the bead capture sequence, alongside a pool and antibody specific barcode for resolution. Conjugates were quantified using Protein Qubit

(Fisher cat. Q33211) for antibody titration and flow validation. In addition, we orthogonally quantified the antibodies using protein BCA (Fisher cat. 23225). For the human donor mixing experiment, CD4 and CD20 antibodies (Biolegend cat. 300541, 302343) were conjugated as described above. Antibodies used in PBMC and comodality experiments are listed in Supplementary Table 1 and 2.

Antibody-specific hybridization design

After conjugation of primary handle oligos, antibodies were combined and pools of oligos (IDT) were used to hybridize the primary handle sequences prior to staining. Of note, only one conjugation was done per antibody with the previously mentioned 20 bp oligonucleotide (e.g. all CD4 conjugates have the same 20 bp oligonucleotide). To avoid non-specific transfer of oligonucleotides between the different antibody clones and the same antibody clone from different wells, each clone received a unique 20 bp handle (Antibody handle). To sequence with antibody and pool specificity, a 10 bp barcode was added to the secondary oligo. The total oligonucleotide sequence consisted of a reverse complementary sequence to the antibody specific primary handle sequence (20 bp), TruSeq Read2 (34 bp), pool barcode (10 bp), and capture sequence (22 bp) (Fig. 1b). Prior to cell staining, 1 μg of each antibody was pooled and hybridized with 1 ul of respective secondary oligonucleotides at 1 μM at room temperature for 15 minutes. The hybridized antibody-oligonucleotide conjugates were purified using an Amicon 50K MWCO column (Millipore cat. UFC505096) according to the manufacturer's instructions to remove excess free oligonucleotides.

Determination of non-specific transfer of oligonucleotides between antibodies

To determine the optimal concentration of hybridizing oligonucleotides for cell staining, we performed a mixed cell line experiment to determine the level of background staining of free oligonucleotides. A mixture of lymphoblastoid cells and primary monocytes were stained with CD14 and CD20 antibodies and hybridized with oligonucleotides with different fluorophores (FAM and Cy5 respectively) per antibody for 15 minutes at room temperature. Concentrations of hybridizing oligonucleotides with different concentrations (1μM and 100 μM) were tested (Extended Data Fig. 10a). Antibodies directly conjugated to fluorophores served as a positive control antibodies (CD13-BV421, Biolegend cat. 562596) to gate respective populations.

Validation of saturation of hybridization oligonucleotides using flow cytometry

To determine the saturation of available primary oligo handles, 1 μg of conjugated CD3 antibody (Biolegend cat. 300443) was hybridized with 1 μl of 1 μM of a reverse complementary oligo with a Cy5 modification (IDT modification /5Cy5/). After incubating at room temperature for 15 minutes, 1 μl of 1 μM of the same reverse complementary oligo but with a FAM modification (IDT modification /56-FAM/) was added to the reaction and additionally incubated for 15 minutes. The cocktail was then added to 1×10^6 PBMCs pre-stained with Trustain FcX (Biolegend cat. 422302) (Extended Data Fig. 10b).

10x Genomics run for SCITO-seq

Washed and filtered cells were loaded into 10x Genomics V3 Single-Cell 3' Feature Barcoding technology for Cell Surface Proteins workflow and processed according to the manufacturer's protocol. After index PCR and final elution, all samples were run on the Agilent TapeStation High Sensitivity DNA chip (D5000, Agilent Technologies) to confirm the desired product size. Qubit 3.0 dsDNA HS assay (ThermoFisher Scientific) was used to quantify the final library for sequencing. Libraries were sequenced on a NovaSeq 6000 (Read1 28 cycles, index 8 cycles and Read2 98 cycles). R2 cycle can be reduced further for cost reduction (depending on the number of pool+antibody barcode length).

Dynamic range assessment of SCITO-seq compared to flow cytometry

PBMCs were collected from anonymized healthy donors and were isolated from apheresis residuals by Ficoll gradient. Cells were frozen in 10% DMSO in FBS and stored in a freezing container (Fisher cat. 5100–0001) at −80°C for one day prior to long term storage in liquid nitrogen. PBMCs from one donor was quickly thawed in a 37°C water bath before being slowly diluted with complete RPMI1640 (Fisher cat.61870-036, supplemented with 10% FBS and 1% pen-strep). Cells were then centrifuged at 300xg for 5 minutes at room temperature. After aspiration, cells were resuspended in Biolegend cell staining buffer (Biolegend, cat. 420201) to a concentration of 1.1×10^7 cells/ml and 900 μl of cell suspension (1×10^7 cells) was transferred to two 5 ml staining tubes and blocked with 50 μl of Human TruStain FcX(Biolegend cat. 422301) for 10 minutes on ice. Ten μg of SCITO-seq CD8a antibody conjugates as well as an isotype control SCITO-seq conjugate were hybridized with 10 μl of 1 μM respective secondary oligos and were cleaned up as described. Cleaned and hybridized SCITO-seq antibodies were mixed with an equal amount of CD8a-APC antibody (Biolegend, cat. 301049) and mouse IgG1-AF488 control (Biolegend, cat. 400129) and added to the cells for 45 minutes on ice in the dark. Cells were washed three times with 3.5 ml cell staining buffer prior to final resuspension at 1×10^7 cells/ml prior to pooling. Samples were stained with 1 μg/ml PI immediately prior to FACS to discriminate live/dead cells (Extended Data Fig. 10c).

Single live cells based on gating were sorted into four bins based on CD8a expression on a BD FACSAriaII. The resulting cells were reanalyzed on the same sorter to determine fluorescent intensities of sorted populations. The remaining cells from each bin were stained with a respective Totalseq-A hashtag (Biolegend, cat. 394601, 394603, 394605, 394607) for 30 minutes on ice washed three times with 3.5 ml cell staining buffer prior to equal pooling and loading 30,000 cells onto a 10× 3'v3 chip.

After libraries were sequenced as mentioned, the resulting fastq files were alligned with Cell Ranger 3.0 Feature Barcoding Analysis with SCITO-seq antibody and hash antibodies specified as an features.csv input. CCDs were filtered manually inspecting the knee plot. Cells from each expression bins were determined based on hashtag counts assigning the maximum count as the corresponding bin. The raw UMI counts for SCITO-seq and scaled fluorescent values for FACS data (from FlowJo (v10)) were log1p transformed for comparison.

Mixed species experiment

HeLa and 4T1 cells were ordered from ATCC (ATCC cat. CCL-2, CRL-2539) and cultured in complete DMEM (Fisher cat. 10566016,10% FBS (Fisher cat. 10083147) and 1% penicillin-streptomycin (Fisher cat. 15140122)) in a 37°C incubator with 5% CO2 on 10 cm culture dishes (Corning). Prior to staining, cells were trypsinized at 37°C for 5 minutes using 1 ml Trypsin-EDTA (Fisher cat. 25200056) and were quenched with 10 ml complete DMEM. Cells were harvested and centrifuged at 300xg for 5 minutes. Cells were resuspended in staining buffer (0.01% Tween-20, 2% BSA in PBS) and counted for concentration and viability using a Countess II (Fisher cat. AMQAX1000). HeLa and 4T1 cells were then mixed at equally and 10^6 cells were aliquoted into two 5 ml FACS tubes (Falcon cat. 352052) and volume normalized to 85 μl. Cells were stained with 5 μl of Trustain FcX for 10 minutes on ice. Cell mixtures were stained with a pool of human and mouse CD29 antibodies, either with the direct or universal design, in a total of 100 μl for 45 minutes on ice. Cells were then washed 3 times with 2 ml staining buffer and centrifuged at 300xg for 5 minutes to aspirate supernatant. Cells were then resuspended in 200 μl of staining buffer and counted for concentration and viability as before. Cells from each stained pooled were mixed and 2×10^4 or 10^5 cells were loaded into the 10x chromium controller using 3'v3 chemistry.

Human donor mixing experiment

PBMCs were prepared as above and resuspended in EasySep Buffer (STEMCELL cat. 20144) at a concentration of 5×10^7 cells/ml before being subject to CD4 and CD20 negative isolation (STEMCELL cat. 17952, 17954). Isolated cells were counted and mixed at a ratio of 3 CD4:1 CD20 for donor 1 and a ratio of 1 CD4:3 CD20 for donor 2 for a total of 1.2×10^6 cells per donor. The cells were centrifuged at 300xg for 5 minutes at room temperature and resuspended in 85 μl of staining buffer and incubated with 5 μl of Human TruStain FcX for 10 minutes on ice in 5 ml FACS tubes. Cells from each donor were either mixed prior or stained with pool specific barcode hybridized antibody oligo conjugates for 30 minutes on ice. Staining was quenched with the addition of 2 ml staining buffer and washed as previously mentioned. Cells were resuspended in 0.04% BSA in PBS and cells from each well were counted, pooled equally, and then passed through a 40 μm strainer (Scienceware cat. H13680–0040). The final strained pool was counted once more prior to loading into a 10x chip B with 2×10^5 cells.

Mass cytometry of healthy controls

PBMCs were isolated, cryopreserved, and thawed from the same donors as previously described. Once thawed, the cells were counted and 2×10^6 cells from each donor were aliquoted into cluster tubes (Corning cat. CLS4401–960EA). Cells were live/dead stained with cisplatin (Sigma cat. P4394) at a final concentration of 5 μM for 5 minutes at room temperature. The live/dead stain was quenched and washed with autoMACS Running Buffer (Miltenyi Biotec cat. 130-091-221). Cells were then stained with 5 μl of TruStain FcX for 10 minutes on ice before surface staining. Mass cytometry antibodies were previously titrated using biological controls to achieve optimal signal to noise ratios. The antibodies in the panel were combined into a master cocktail and incubated with cells from the two donors

and stained for 30 minutes at 4°C. After washing twice with 1 ml autoMACS Running Buffer, the cells were resuspended and fixed in 1.6% PFA (EMS cat. 15710) in MaxPar PBS (Fluidigm cat. 201058) for 10 minutes at room temperature with gentle agitation on an orbital shaker. Samples were then washed twice in autoMACs Running Buffer, and then three times with 1X MaxPar Barcode Perm Buffer (Fluidigm cat. 201057). Each sample was then stained with a unique combination of three purified Palladium isotopes obtained from Matthew Spitzer and the UCSF Flow Cytometry Core for 20 minutes at room temperature with agitation as previously described[1]. After three washes with autoMACS Running Buffer, samples were combined into one tube and stained with a dilution of 500 μM Cell-ID Intercalator (Fluidigm cat. 201057), to a final concentration of 300 nM in 1.6% PFA in MaxPar PBS at 4°C until data collection on the CyTOF three days later. Immediately before running on the CyTOF machine, the sample tube was washed once with each 15 ml autoMACS Running Buffer, MaxPar PBS, and MilliQ H2O. Once all excess proteins and salts were washed out, the sample was diluted in Four Element EQ Calibration Beads (Fluidigm cat. 201078) and MilliQ H2O to a concentration of 10^6 cells/mL and run on a CyTOF Helios at the UCSF Flow Cytometry Core.

Comparing Mass cytometry (CyTOF) and SCITO-seq

Data was normalized and de-barcoded using the premessa package (https://github.com/ParkerICI/premessa). Normalized and demultiplexed files were uploaded to Cytobank (https://www.ucsf.cytobank.org/) for gating and manual identification of immune cell subsets. Files containing only singlet events were exported from Cytobank and analyzed with CyTOFKit2 package (https://github.com/JinmiaoChenLab/cytofkit2). Through CyTOFkit2, events were clustered using Rphenograph with k=150 and visualized via UMAP to calculate cell type proportions. The cost breakdown of SCITO-seq and comparison with CyTOF is in Supplementary Tables 3 and 4.

Pre-processing and initial filtering

Both the species mixing experiments and human donor mixing experiments were processed using Cell Ranger 3.0 Feature Barcoding Analysis with default parameters. For cDNA and ADT alignment, we specified the input library type as ‘Gene Expression’ and ‘Antibody Capture’ respectively. For ADT alignment, specific barcode sequences (Ab+pool) were specified as a reference. Reads were aligned to the hg19 and mm10 concatenation reference the species mixing experiments. For human experiments, the reads were aligned to the human reference genome (GRCh38/hg20).

Normalization for species mixing and T/B cell human donor mixing experiment

For transcriptomic counts, data was normalized by dividing each cell’s UMI counts to the total UMI counts and multiplied by 10,000. Then, the data was log1p transformed (numpy.log1p). Finally, the data was scaled to have mean = 0 and standard deviation = 1. Clustering was done using the Leiden algorithm[2] using 10 nearest neighbors at a resolution of 0.2 for mixed species experiments and two-donor experiments with two cell types (T and B cells).

To normalize ADT counts in species mixing experiment, the data was log transformed and standardized to have mean = 0 and standard deviation = 1. For ADT counts in two human donor mixing experiment with two cell types, after log transformation of the raw data, we used a Gaussian Mixture Model in scikit-learn package in python to normalize the data with the following parameters (convergence threshold 1e-3 and max iteration to 100, number of components 2). The data was normalized by a z-score like transformation (log transformed raw value - mean of the posterior means of two components / mean of the posterior standard deviations).

Implementation of an algorithm for multiplet resolution

Considering all antibodies in each pool, we normalized each value by dividing mean expression value of CD45 counts across all pool (considered as a universal expression marker on PBMCs) for each droplet barcode yielding a $p \times m$ matrix (where p is the number of pool and m is the number of droplet barcodes). Then, the matrix was CLR normalized and resolved using HTODemux from Seurat (v3.0) (http://satijalab.org/seurat/) to classify the droplet barcode to a pool or unassigned (we discretized the value of 0 or 1). Using this binary matrix, we iterated over p times (where discretized value equals 1) to get final resolved matrix of ($n \times r$) where n is the number of antibodies used and r is the resolved number of cells. For each iteration, we selected the columns that were positive for the above-mentioned discretized matrix. An additional round of HTODemux was used to re-classify the 'Negative' cells from initial classification because most of the initial classification which deemed the cells negative had a UMAP distributions which were contained in the original clusters.

Analysis of PBMC experiment

Normalization and resolution of multiplets—To normalize transcriptomic data for our PBMC experiments, we used the same normalization method as described above. To generate the UMAP based on ADT counts for the PBMC experiment, we performed pool demultiplexing using the algorithm described previously. Then, the resolved matrix ($n \times r$) was normalized as in the cDNA processing. Raw values were normalized to total counts of 10,000 per cell and log1p transformed. Then, the values were standardized (mean 0, standard deviation 1) per pool. Using these normalized values, PCA was performed to reduce dimensionality. Leiden clustering was done with 10 neighbors and 15 PCs from the previous step. A resolution value of 1.0 was used to assign clusters for the whole PBMC experiments. Finally, UMAP was utilized to visualize the resolved total cells. To remove collided cells in 60-plex and 165-plex experiment, we computed the average number of UMIs expressed per cell and thresholded cells based on the quantile distribution (>80% in the UMI distribution is filtered out) to remove cells and also manually inspect expression across all Leiden clusters to exclude the cluster that expresses multiple markers.

Demultiplexing donor identity—For demultiplexing donors, a VCF file containing donor genotype information and the bam file output from the Cell Ranger pipeline were used as inputs for demuxlet (Freemuxlet) with default parameters. For donors without genotypic information, we used Freemuxlet (https://github.com/statgen/popscle/) to assign droplet barcodes to a corresponding donor.

SCITO-seq with the 10x ATAC-seq kit—We initially designed a secondary oligo compatible with the 10x ATAC-seq kit by changing the hybridizing end of the splint oligo to the reverse complement of the Read 1 Nextera sequence) from the feature barcode capture sequence (10x 3'v3). We modified the microfluidic cell and enzyme mixture to the following mastermix; 4 μl of 10mM dNTP, 16 μl of RT buffer (5x), 4 μl of Maxima H minus, and cells and RNAse free water up to 80 μl. After running the solution through a 10x chip E reaction as in the 10x user guide, the GEMs were thermocycled at 53C for 45 min and 85C for 5 min. The emulsion was broken as in the 10x user guide and ADT fragments were eluted in 40 μl. We performed an index PCR with the following conditions: 40 μl of sample, 50 μl of 2x KAPA HiFi HotStart ReadyMix, 1 μl each of P5 primer (100 uM) and universal read 2 Nextera primer, and 8 μl of RNAse-free water. The sample was cycled as follows: initial denaturation at 98C for 45s, cycled 12x at 98C for 20s, 54C for 30s, and 72C for 20s, followed by a final extension at 72C for 1 min.

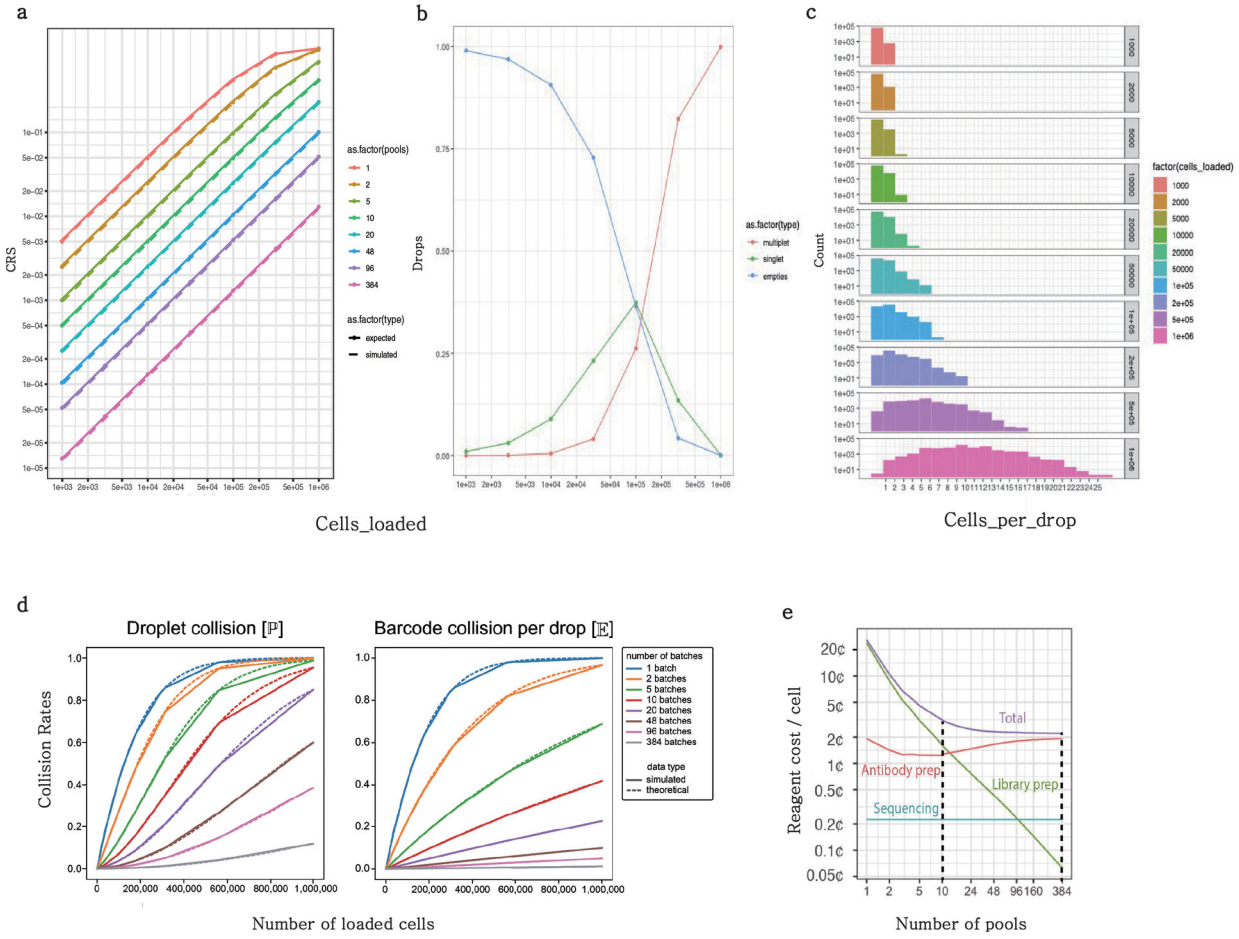
SCITO-seq with a commercial antibody panel—To scale SCITO-seq to a commerical platform, we modified our secondary oligo to be compatible with Biolegend's TS-C platform (normally used for the 10× 5' kits) for the 10× 3'V3 kit. To do this, we changed the antibody hybridization region in our original 3'v3 design to the reverse complement of antibody specific TS-C barcode (15bp) sequences (Extended Data Fig. 7a). After emulsion breakage, we followed the index PCR protocol as per manufacturer's recommendations (10x Genomics, CG000185 Rev D, page 52).

Comodality experiment—To generate compatible secondary oligos with scifi-RNA-seq, we conjugated unique 20 bp 5' amine modified oligos to each of our six antibodies, varying from our previous 3' amine conjugation to present a favorable orientation of the secondary oligonucleotide for capture in a similar fashion to transcripts in the scifi-RNA-seq workflow. In addition, we spiked-in an additional orthogonal bridge oligo for the in-emulsion ligation to reduce competition of transcripts and ADT molecules for the bridge oligo (Extended Data Fig. 9b). We stained 5 pools of a mixture of 5 cell lines (cat #; NK92: CRL-2407™, HeLa: CCL-2™, Jurkat, Clone E6–1: TIB-152™, 4T1: CRL-2539™, LCL: CRL-1855™) for 30 min prior to washing and executing the scifi-RNA-seq protocol. After the scifi-RNA-seq workflow, we loaded 3×10^4 into the 10x chromium controller using the 10x ATAC-seq kit. After emulsion breakage as in the 10x user guide, we saved 4 μl of the 24 μl silane bead elution for ADT library construction. The ADT sample index PCR reaction was set up with 4 μl of sample, 5 μl of P5 primer (10 μM), 5 μl of i7 index primer (10 μM), 50 μl of KAPA HiFi mastermix, and 36 μl of RNAse-free water. Cycling conditions were as follows: 98C for 45s, followed by 12 cycles of 98C for 20s, 54C for 30s, 72C for 20s, and ending with a final extension of 72C for 1 min. We cleaned up and selected the fragments using AMPure XP beads at a ratio of 1.2X, prior to a final elution in 20 μl. To construct the gene expression library, we used a plexWell 96 Library Preparation kit (Seqwell ref PW096–1) to tagment 10 ng of DNA per reaction. This pre-loaded Tn5 was used to ease the number of tagmentations in the scifi-RNA-seq workflow and increase the reproducibility with a commerical product over custom-loaded Tn5s. The final gene expression library sample index PCR was performed as-is in the scifi-RNA-seq workflow. The resulting libraries were

sequenced on a Novaseq 6000 S1 v1.0 flow cell with the following read configuration: 21:8:16:78 (Read1:i7:i5:Read2)

Comodality data processing—To process the transcriptomic data, the generated fastqs (R1:21bp, R2:16bp, R3:78bp) were stitched to make a final R1 file containing a droplet barcode (16bp) + well barcode (11bp) + UMI (8bp) per read. We used kallisto version 0.46.1 and specified the cell barcode as 27 bp (16+11; droplet and well barcode bp lengths) and ran bustools to produce count matrices (https://www.kallistobus.tools/getting_started). To process the ADT fastqs (same read configuration as RNA) were stitched to produce a final R1 file (35bp), R3 data was trimmed to 10bp (encoding antibody barcode) for barcode alignment. These reads were then processed using a modified dropseq pipeline (v2.4.0; aligner swapped to bowtie (v2.4.2)) (https://github.com/broadinstitute/Drop-seq/releases). Counts were then normalized as done in the PBMC experiment above for both ADT and RNA. RNA genes were determined based on manual curation after running the Wilcoxon's test for determining highly variable marker genes. For overlap analysis in Fig 4g, gene scores (using scanpy's function) for each cell lines are calculated and standardized (mean:0, variance:1, z-score to represent the classification accuracy) to be used as an input for the heatmap generation (Seaborn package's (v0.11.1) heatmap function).

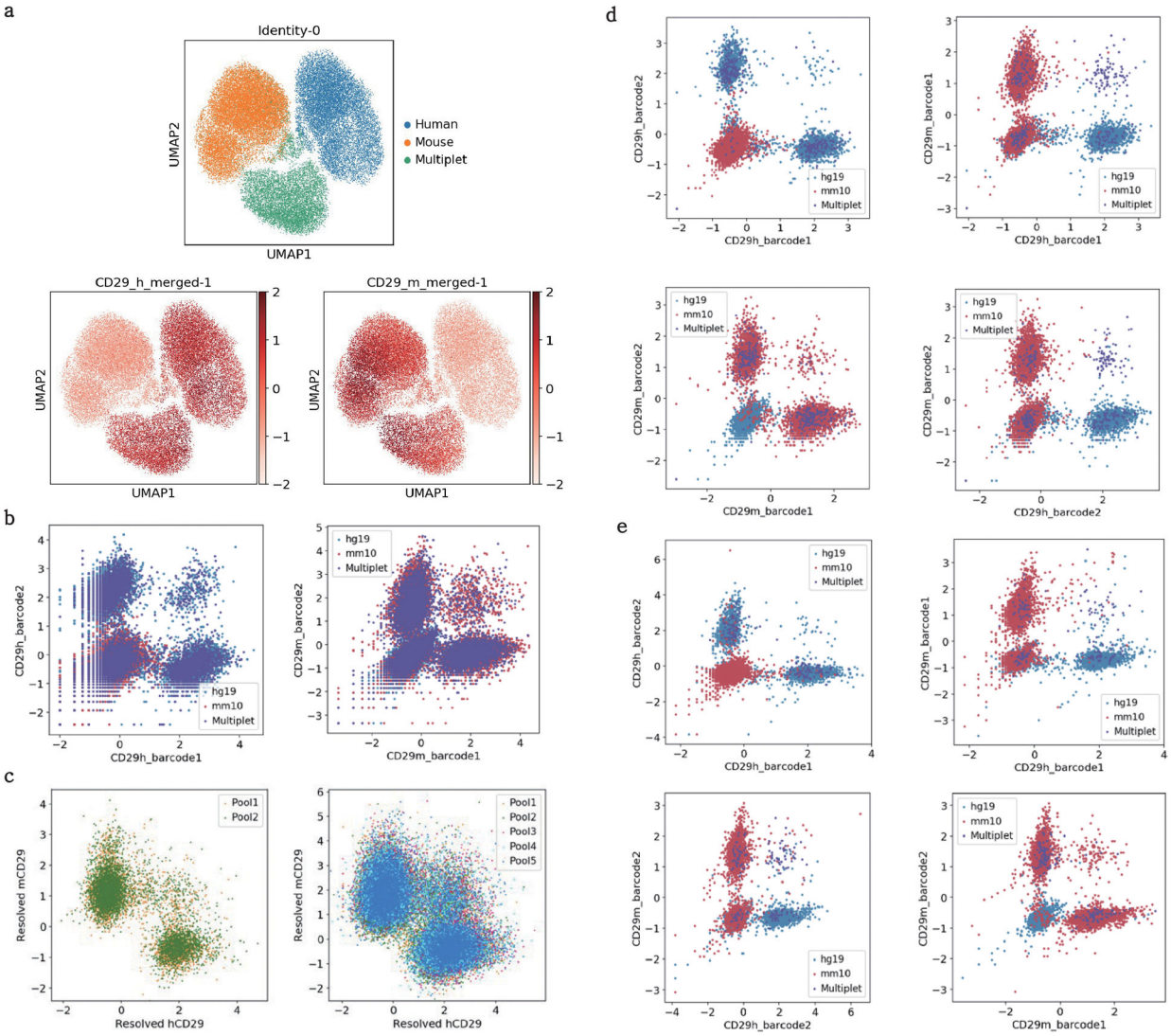
Extended Data



Extended Data Fig. 1. Simulation and cost analysis of SCITO-seq

a, Collision rate (y-axis, denoted as CRS) as a function of the number of cells loaded and the number of pools (denoted by different colors). Here, number of simulations were performed as follows: maxSim = 1000 for cells_loaded <= 1e4, maxSim = 100 for 1e4 < cells_loaded <= 1e5, and maxSim = 10 for 1e5 < cells_loaded. Expected (solid lines) and simulated (dotted lines) collision rates are based on the Poisson statistics for 100,000 droplets and the number of droplets containing cells is modeled as 0.6 × 100,000 in simulation. When the number of cells loaded is not large (e.g., less than 10,000), there is noticeable variance in the number of collisions, so multiple simulation runs were used to estimate the collision rates shown in dotted lines. **b**, Number of droplets (y-axis) containing no cells (blue), exactly one cell (green) or greater than one cell (red) as a function of the number of cells loaded (x-axis). Singlets refer droplets that contain one cells, multiplets contain more than one cell (>2) and empties meaning no cells in the droplet. **c**, Distribution (Y-axis, counts) of number of cells per droplet (X-axis) for different cell loading numbers (cells_loaded) based on Poisson distribution. **d**, (Left) Droplet collision rate, depicting proportion of droplets with at least one barcode collision. (right) Barcode collision rate, estimating proportion of batches (pools) with a collision in a given droplet. Collision rates were calculated using simulations of a Poisson Point Process (solid lines) or a closed form solution (dashed

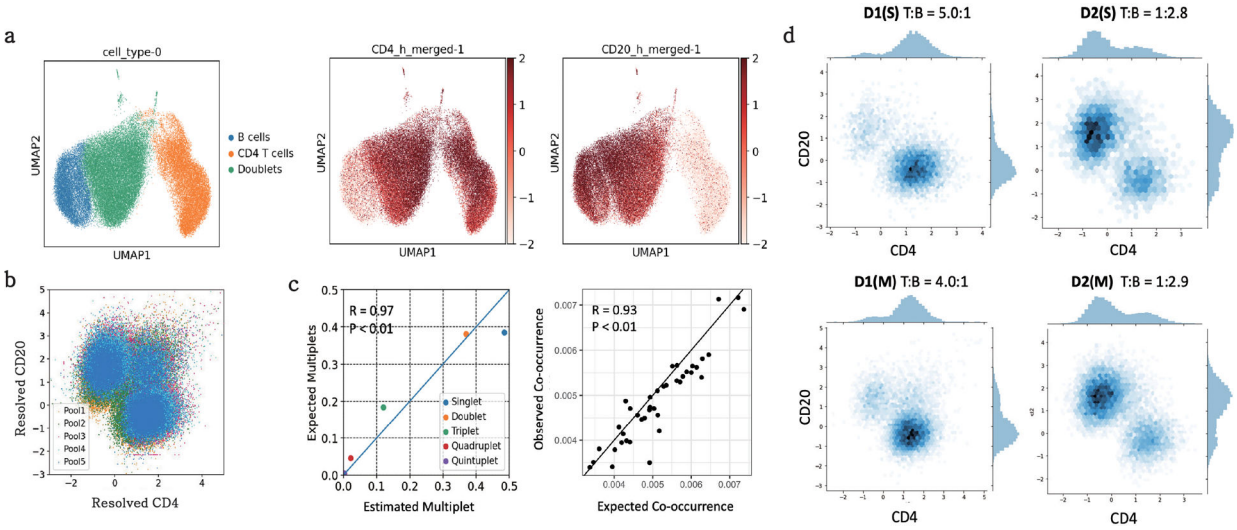
lines; see Methods). Estimates from a closed form solution robustly and almost identically recapitulate simulations and can be used to calculate collision rates for an experiment. **e**, Total cost estimates (purple) including library prep (green), antibody prep (red) and sequencing cost (blue) assuming 40 reads/Ab/cell and a panel of 30 antibodies for different number of SCITO-seq pools.



Extended Data Fig. 2. Species mixing QC analysis (Human/Mouse)

a, (Upper) Transcriptomic UMAP of human and mouse cells, distinguished by transcript alignment. (Lower) ADT staining of mouse and human cells overlaid on the transcriptomic UMAP for 100k loading experiment. Pool barcodes per antibody were merged (i.e. CD29_h_merged-1 = CD29_h_barcode-1+….+CD29_h_barcode-5, where the latter number represents the pool number). Species classification was transcriptomically determined by a >95% cutoff based on normalized counts specific to either species. Cells which did not meet the threshold were classified as Multiplets. Overlaid normalized ADT counts shows human and mouse antibody staining. **b**, Scatterplot showing within species multiplets

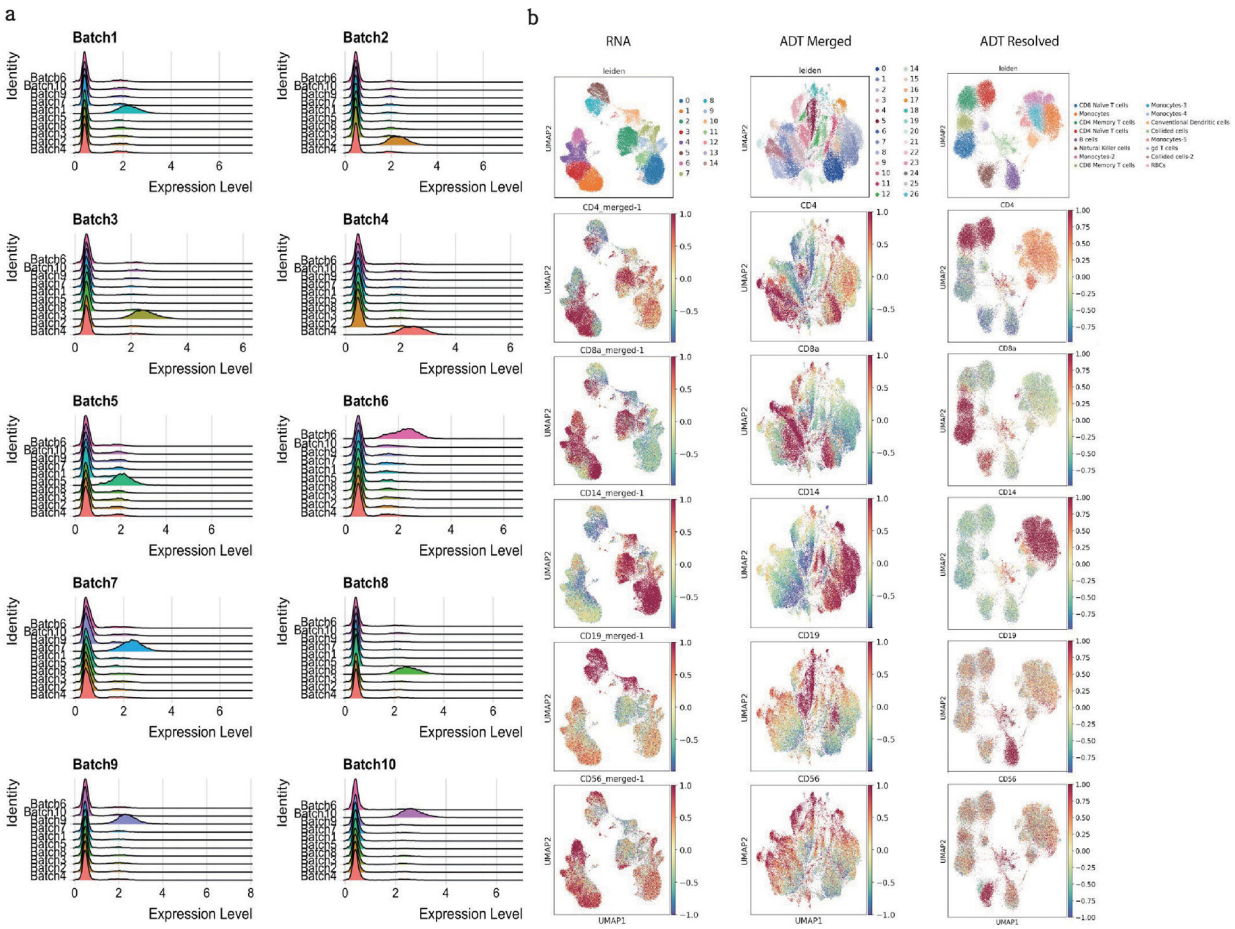
(shown on double-positive axes) across batches when loading 100k cells. Resolution of cell types with a single batch barcode and annotation of Multiplets (positive for both pool barcodes, such as CD29h_barcode1 and CD29h_barcode2 positive or CD29m_barcode1 and CD29m_barcode2). **c**, Scatterplots for species mixing 20k (left) and 100k (right) loading experiments colored by pool showing pool specific staining level. Resolved hCD29 or mCD29 on the axes refers to normalized antibody counts after resolution into single cells. If a droplet contained a mixture of hCD29 from pool 1 and pool 2, the droplet was resolved as two cells with the pool-normalized counts. **d**, Scatter plots of SCITO-seq normalized counts from 2×10^4 loading of species mixing to determine cross pool or within pool background level. **e**, Scatter plots of directly-conjugated hCD29 and mCD29 antibody-based normalized counts from 2×10^4 loading of species mixing to show cross pool or within pool background level using direct conjugation. Hg19, mm10, and Multiplet define cell populations based on respective transcriptomic alignment. Direct conjugates provide a baseline for noise in the SCITO-seq system.



Extended Data Fig. 3. Species mixing QC analysis (Human B/T cells)

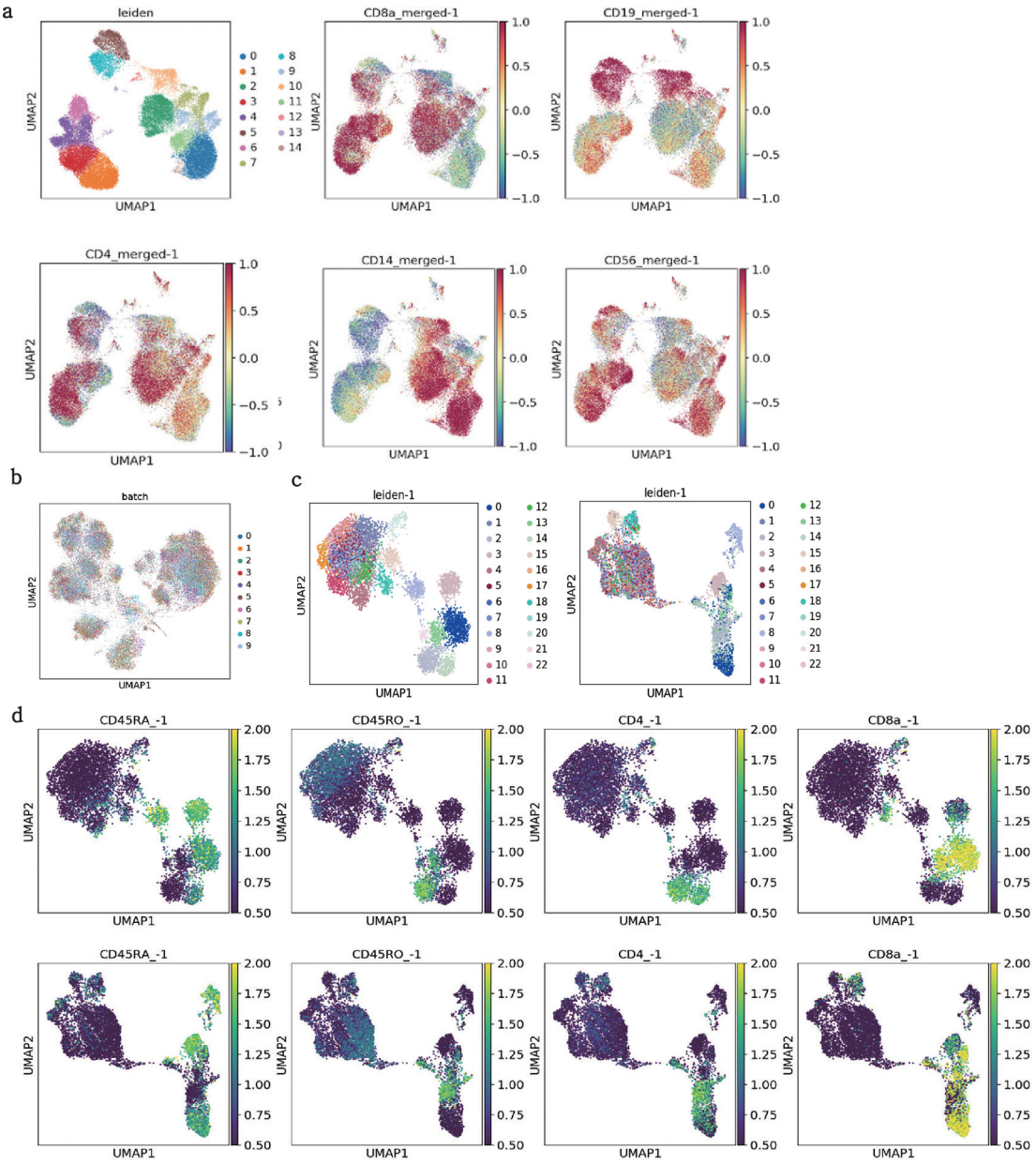
a, UMAP projection (left) of T/B cell experiments with 2×10^5 loading colored by cell types as determined transcriptomically (cutoff value of 0.9 for differences in highly variable genes). Doublets (multiplets) represent a mixture of T and B and are colored in green. The other two panels demonstrate specific staining of merged (merging all 5 pool barcodes) and normalized SCITO-seq counts. **b,** Scatterplot for 200k T/B experiments loading colored by pool. Resolved hCD4 or hCD20 on the axes refers the normalized antibody count after resolution into single cells. For example, if droplet is a mixture of hCD20 from pool 1 to 5, the resolved count should be either of the normalized counts from specific pool only (for Pool1 legend, the Resolved axes are represented by the normalized pool 1 counts). **c**, Estimated (x-axis) versus expected (y-axis) frequencies of Multiplets (frequencies of droplets that contain 1 cell to 5 cells) between estimated (observed) vs expected (simulated) for 2×10^5 loading experiment (left). The five dots represent the number of the cells in the droplets (from single to five cells). Expected (x-axis) versus observed (y-axis) frequencies (right) of co-occurrences between antibody pool barcodes for loading concentrations of

2×10^5 cells. Expected frequencies were calculated based on the frequencies of barcodes in singlets. Correlation value and p-value are also shown. Observed co-occurrence of antibody pool barcodes is calculated using R package mixtools (v1.0.4) implementing normalmixEM function with default parameters (epsilon = 1e-08, maxit = 1000). **d**, Distribution of the normalized UMI counts for each antibody in cells resolved from droplet containing single pool barcodes (S) and multiple pool barcodes (M) per donor for 200k loading experiment. Distribution of the antibodies in pool multiplets shows expected prior mixture proportions (5:1 for donor1 (D1) and 1:3 for donor2 (D2)) and overlaps with the corresponding distribution in pool singlets



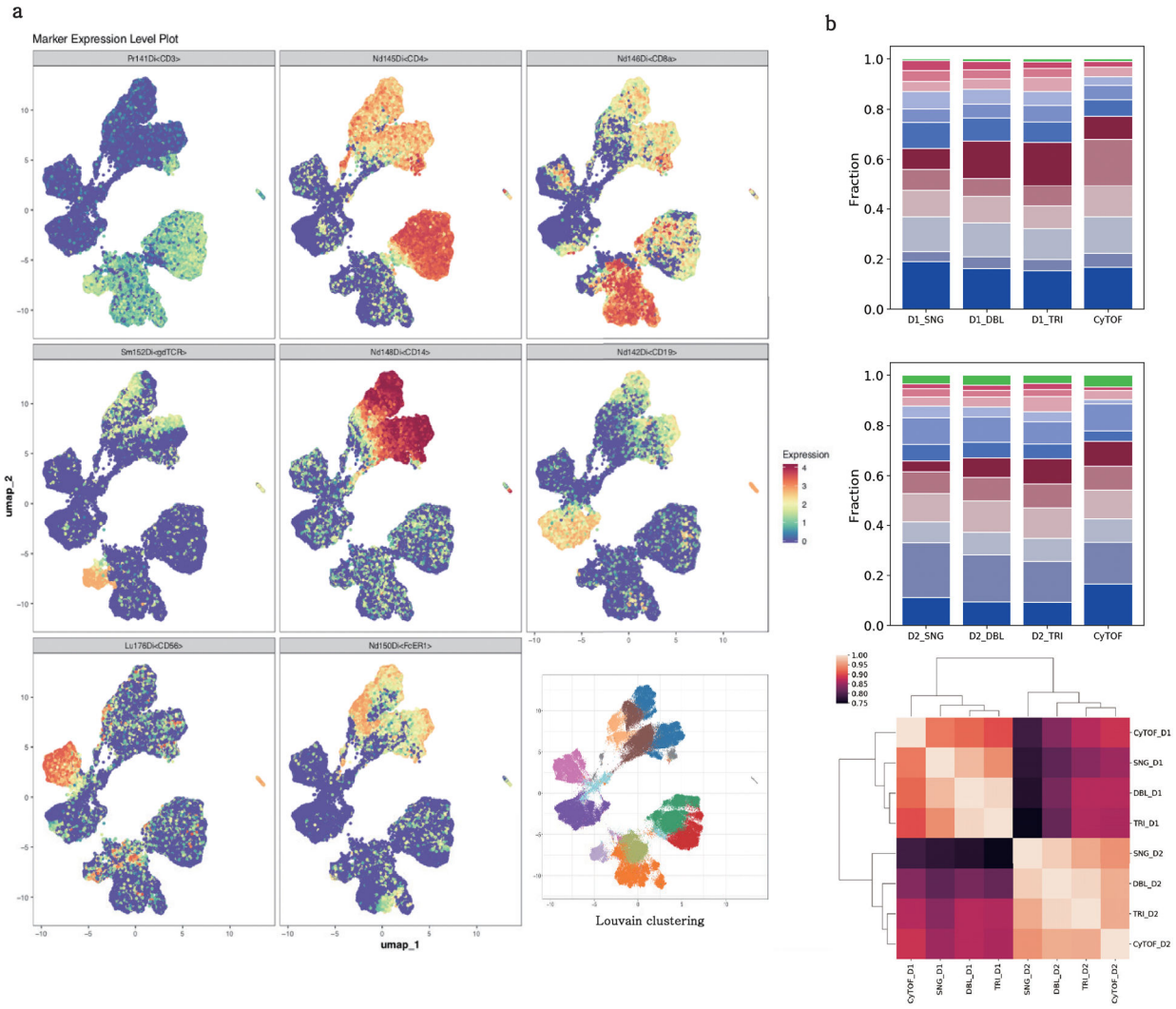
Extended Data Fig. 4. Human donor PBMC QC analysis (28plex)

a, Ridgeplots of 28-plex experiment of the pool specific antibodies. The normalized expression values of 60 antibodies within each pool was summed for thresholding. An individual plot contains the batch specific normalized expression values for demonstrating the signal to noise distribution of the expected specificity. **b**, UMAP projections for 100k PBMC loading experiment for representative markers. UMAP projection comparison using RNA expression (left), merged antibody counts before resolution (middle) and after resolution (right). The scale represents the normalized merged ADT counts (left and the middle) and resolved ADT counts (right).



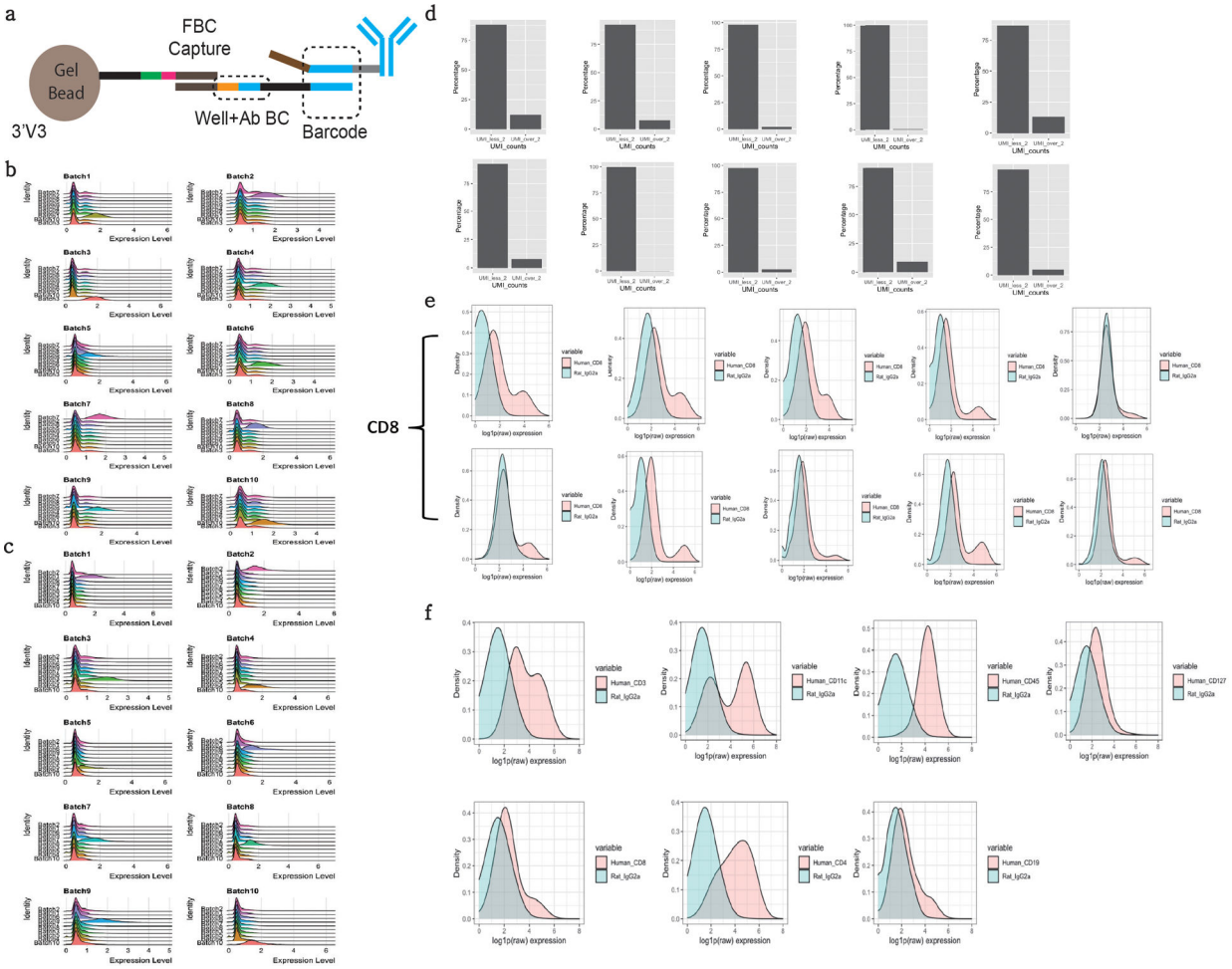
Extended Data Fig. 5. Human donor PBMC QC analysis 2

a, RNA expression based UMAP projections for representative markers of 200k PBMC loading. Since the RNA molecules are not combinatorially indexed, these UMAPs show stark contrast with the resolved UMAP based on normalized ADT counts where we see clear distinction of all clusters. **b**, UMAP ADT projection of 200k loading PBMC data colored by different pools (color numbers 0 through 9). Two pooled donors prior to aliquoting into 10 different pools to investigate batch effects across all stained wells and found no significant batch effects. **c**, ADT UMAP clusters overlaid on ADT UMAP (left) and ADT UMAP clusters overlaid on transcriptomic UMAP. **d**, (top) Protein expression on ADT UMAP of CD4/8 and CD45RA/RO. (down) Protein expression on transcriptomic UMAP of CD4/8 and CD45RA/RO.



Extended Data Fig. 6. Human donor PBMC QC analysis 3

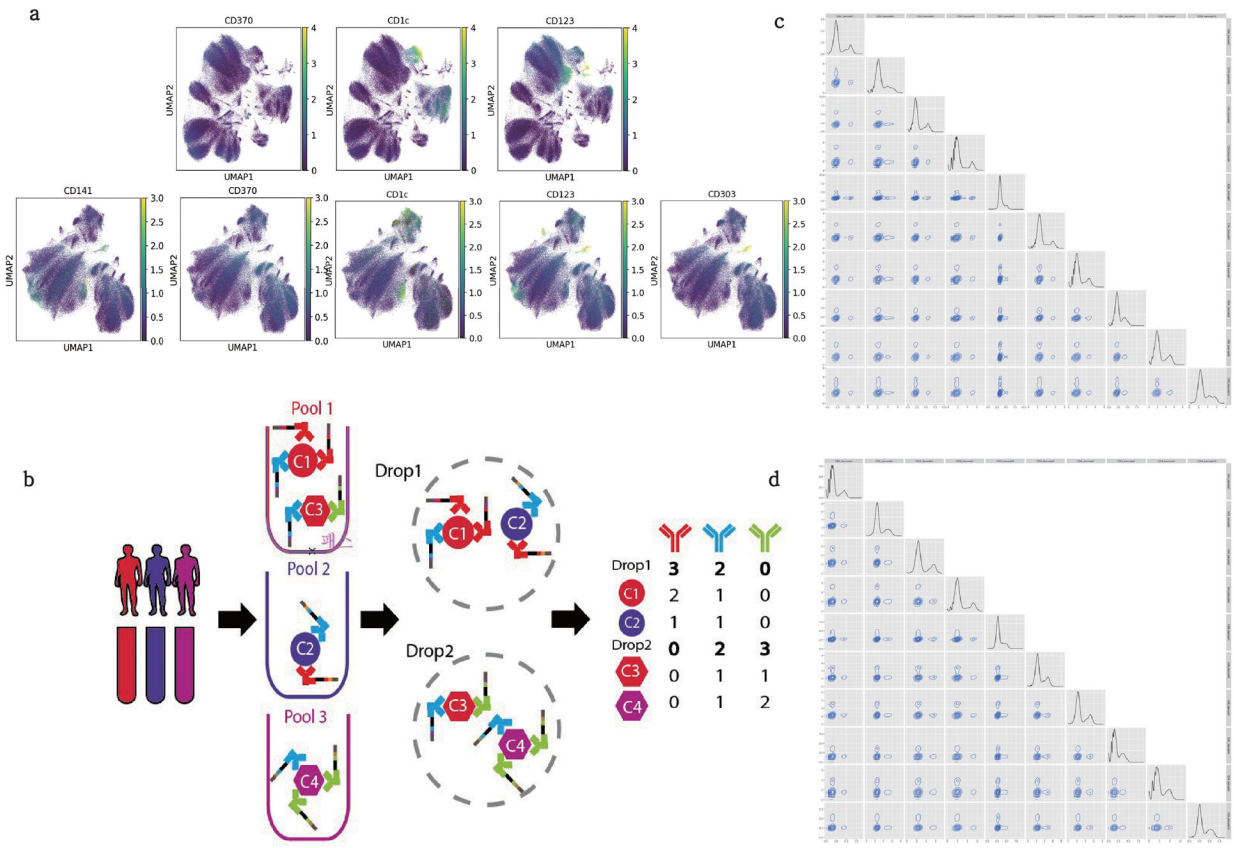
a, UMAP (x-, y- axis with UMAP1 and UMAP2 dimensions) with representative PBMC markers based on CyTOF experiment using the same donor and antibody panel as in SCITO-seq. The scale shows arcsinh (hyperbolic inverse sine) transformed normalized values. **b**, Comparisons of SCITO-seq with CyTOF per donor (D1: top, D2: middle) for 100k loading data (SNG:singlet, DBL:doublet, TRI:triplet). Pairwise correlation heatmap plot (bottom) is also shown (similar to Fig 2d and e in the main figure). Within donors, the proportion each Leiden cluster was highly correlated (Cosine similarity within donor1:0.95, donor2:0.94).



Extended Data Fig. 7. Scalability experiment and QC analysis (60- and 165-plex)

a, Design of splint oligo with Totalseq-C compatible system. The splint oligos (FBC RC (reverse complement) + Well + Ab BC (5+5,10bp) + Read2 + Totalseq-C Barcode RC) are hybridized to the barcode region of the Totalseq-C oligo conjugated antibodies (right dotted lines around the blue region). 1uM splint oligo is also used and incubated 15min of hybridization (same workflow as conventional SCITO-seq). The well and the antibody barcode sequences are encoded in orange and blue above. **b,** Ridgeplots of 60-plex experiment showing the specificity of the pool specific antibodies. The normalized expression values of 60 antibodies with 10 pool result is shown above. Individual plot contains the batch specific normalized expression values to show signal to noise distribution of the expected specificity (first Batch1 plot is expected show a shift of Batch1 only (all batch1 60 normalized anitbody counts are aggregated). **c**, Ridgeplots of 165-plex experiment showing the specificity of the pool specific antibodies. The normalized expression values of 60 antibodies with 10 pool result is shown above. Individual plot contains the batch specific normalized expression values to show signal to noise distribution of the expected specificity (first Batch1 plot is expected show a shift of Batch1 only (all batch1 165 normalized anitbody counts are aggregated). **d**, Barplots of 165-plex experiment showing low UMI counts of an example Isotype control ADT counts (Rat IgG) for all 10 pools (top 1–5

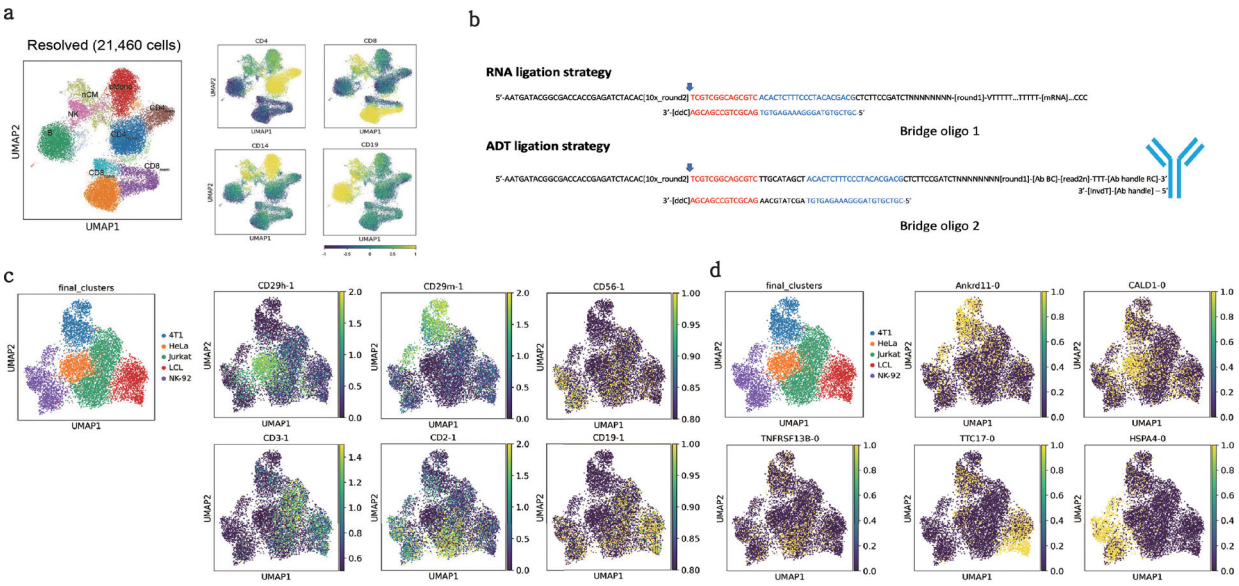
batches, bottom 6–10 batches). The percentage of cells (y-axis) that express the ADT less than 2 UMI or over is calculated. Background noise across batches shows less than 2 UMI counts in ~92% of the cells. **e**, Overlay density histograms of the example CD8 vs Isotype control Ab (Rat IgG) to assess the 'noise' level for all 10 pools (top 1–5 batches, bottom 6–10 batches) in 165-plex data. X-axis for log1p(raw counts) transformed values and y-axis for density. **f**, Overlay density histograms of the example antibodies aggregated over all 10 pools (CD3, CD11c, CD45, CD127, CD8, CD4, CD19) vs Isotype control Ab (Rat IgG) to assess the 'noise' level in 165-plex data. X-axis for log1p(raw counts) transformed values and y-axis for density.



Extended Data Fig. 8. Scalability experiment and QC analysis 2 (60- and 165-plex)

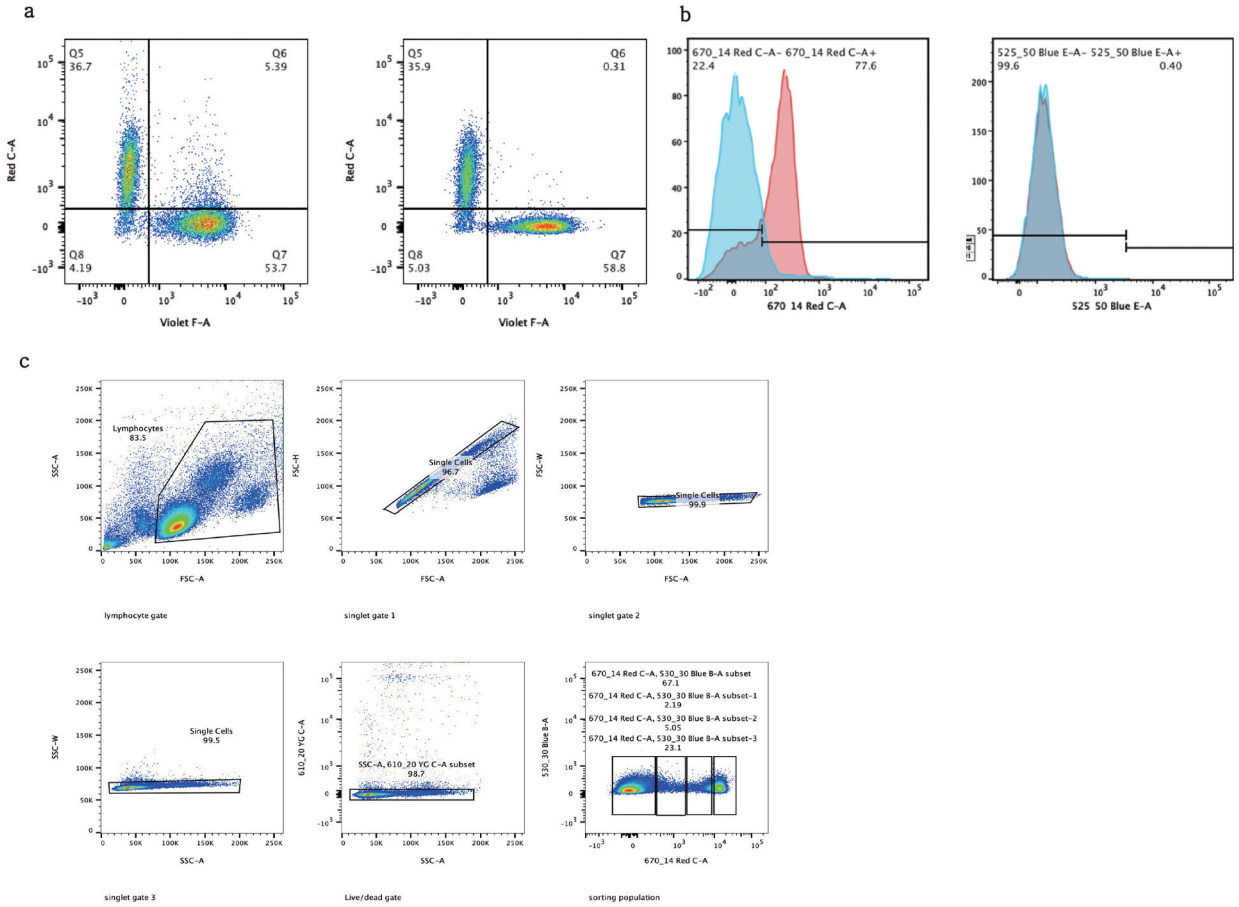
a, UMAPs of 60-plex (upper) and 165-plex (lower) experiment showing normalized expression of cDC1, cDC2 and pDC markers. CD141 and CD370 for cDC1 and CD1c for cDC2 and CD123 and CD303 for pDC markers. **b**, Schematic of sample multiplexed SCITO-seq where different samples are hashed with different pool barcodes (Red, Blue, Purple). Droplets containing cells from different individuals (two different colors) can be resolved into separate cells. **c**, The example of pairwise correlation plots (using the ggpair R package) of normalized expression values of all 45 combinations (combinations of choosing 2 pairs from 10 pools) for CD4 antibody in 60-plex experiment. If spillover is present (if secondary oligo that encodes pool1 is not washed sufficiently, this could hybridize to other conjugated handle (same antibody) from different well), you would expect have staining on the double positive axis, which we do not see in this experiment. **d**, Another example of

pairwise correlation plots (using the ggpair R package) of normalized expression values of all 45 combinations (combinations of choosing 2 pairs from 10 pools) for CD4 antibody in TSC 165-plex experiment. If spillover is present (if secondary oligo that encodes pool1 is not washed sufficiently, this could hybridize to other conjugated handle (same antibody) from different well), you would expect have staining on the double positive axis, which we do not see in this experiment.



Extended Data Fig. 9. Development of comodality experiment and QC analysis

a, Proof-of-concept experiment to analyze SCITO-seq using ATAC-kit. Representative PBMC with 12 surface markers (CD4, 8, 14, 16, 45, 45RA, 45RO, 19, 20, 56, 11c, HLA-DR) are stained in 5 separate pools loading 50k cells in this experiment showing specific staining profiles above (nCM: non-conventional monocytes, cMono: conventional monocytes). **B**, Schematic of the comodality experiment during the GEM ligation step using 10x Genomics ATAC kit. Detailed sequence structure of the RNA and ADT capture during the GEM reaction using the scifi-RNA-seq workflow. A more detailed workflow for the RNA can be found in the Supplementary Figure 2 in the scifi-RNA-seq paper. 10x_round2 refers to the 16bp droplet barcode, round1 barcode refers to the well barcode (11bp) used in the *in-situ* reverse transcription reaction. Untemplated 'CCC' is add at the end of the reverse transcription reaction. Antibody barcode (Ab BC fixed 10bp) and antibody handle (Ab handle fixed 20bp, conjugated directly to the blue antibody) sequences are specific to the antibody. Read2n stands for Read2 Nextera sequence. Compared to the bridge oligo 1 (used to capture in-situ RT mRNA molecules), bridge oligo 2 has extra 10bp (AACGTATCGA between red and blue colored sequences). ddC (dideoxy C) and InvdT (inverted dT) for preventing extension. Arrow indicates the ligation site during the GEM reaction. **c**, Dimensional reduction using UMAP with normalized RNA counts and corresponding cell line specific ADT marker expressions on the UMAP space. **d**, Dimensional reduction using UMAP with normalized RNA counts and corresponding single RNA marker expressions on the UMAP space.



Extended Data Fig. 10. Flow validation experiment of SCITO-seq

a, To reduce the non-specific staining of secondary oligonucleotides, we titrated oligonucleotides at 1uM (right) and 100uM (left). After hybridization of oligonucleotide conjugated antibodies with a Cy5 conjugated reverse complementary oligonucleotide for 15 minutes, a mixture of LCLs and primary monocytes were stained with the hybridized material an CD13-BV421 for 30 minutes, washed twice and analyzed on a LSRII. CD13 BV421 antibody was captured by the Violet-F channel (x-axis) and Cy5 tagged secondary oligonucleotides was captured on the Red-C channel to check the level of background staining (Q6 gated population refers to the spillover of non-cognate secondary oligonucleotides in the primary monocyte population). **b**, To determine if 1 ul of 1 uM reverse complementary oligonucleotide would saturate 1 ug antibody, we first hybridized 1 ug of oligonucleotide conjugated CD3 with 1 ul of 1 uM reverse complementary oligonucleotide conjugated to Cy5. Following this, another 1 ul of 1 uM reverse complementary oligonucleotide was added, but with a FAM conjugated instead. This was incubated for 15 minutes before being added to the whole PBMC and washed twice before running on an LSRII. Left figure shows the positive shift (red) when first hybridization occurs, and second histograms shows there is essentially no significant shift because of the near saturation of the first handle sequence. **c**, Lymphocytes were gated for singlets and live cells (Live/dead gate, YG C-A is the PI dye channel) prior to binning

samples across CD8a expression for sorting. Red-C represents CD8a-APC and Blue-B represents isotype control-AF488.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

C.J.Y.,Y.S.S. and M.H.S. are Chan Zuckerberg Biohub Intercampus Research Award Investigators and C.J.Y. and M.H.S. are members of the PICI. C.J.Y. is further supported by the NIH grants R01AR071522, R01AI136972, R01HG011239. E.D.C is supported by UCSF PBBR grant for Center for Advanced Technology. This work was supported by NIH grant DP5 OD023056 and funding from the UCSF PBBR to M.H.S. and NIH grant S10 1S10OD018040, which enabled the procurement of the Helios mass cytometer used in this study. We acknowledge the PFCC (RRID:SCR018206) supported in part by Grant NIH P30 DK063720 and by the NIH S10 Instrumentation Grant S10 1S10OD021822-01. This study was also supported by NIH grant R35-GM134922.

Competing interests

C.J.Y. is a SAB member for and hold equity in Related Sciences and ImmunAI, a consultant for and hold equity in Maze Therapeutics, and a consultant for Trex Bio. C.J.Y. has received research support from Chan Zuckerberg Initiative, Chan Zuckerberg Biohub, and Genentech.

Data availability

Single cell sequencing data generated in this project have been deposited to the Gene Expression Omnibus (GEO) with the accession code GSE147808. Processed data is also available at the website (https://github.com/yelabucsf/SCITO-seq) with tutorials.

MAIN REFERENCES

1. Macosko EZ et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]
2. Klein AM et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015). [PubMed: 26000487]
3. Buenrostro JD et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015). [PubMed: 26083756]
4. Stoeckius M et al. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14, 865–868 (2017). [PubMed: 28759029]
5. Shahi P, Kim SC, Haliburton JR, Gartner ZJ & Abate AR Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci. Rep 7, 44447 (2017). [PubMed: 28290550]
6. Gerlach JP et al. Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. Sci. Rep 9, 1469 (2019). [PubMed: 30728416]
7. Peterson VM et al. Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol 35, 936–939 (2017). [PubMed: 28854175]
8. Bandura DR et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal. Chem 81, 6813–6822 (2009). [PubMed: 19601617]
9. Spitzer MH & Nolan GP Mass Cytometry: Single Cells, Many Features. Cell 165, 780–791 (2016). [PubMed: 27153492]
10. Swanson E et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. eLife vol. 10 (2021).

11. Mimitou EP et al. Scalable, multimodal profiling of chromatin accessibility and protein levels in single cells. doi:10.1101/2020.09.08.286914.
12. Kang HM et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol 36, 89–94 (2018). [PubMed: 29227470]
13. McGinnis CS et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods vol. 16 619–626 (2019). [PubMed: 31209384]
14. Stoeckius M et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 19, 224 (2018). [PubMed: 30567574]
15. Datlinger P et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods 14, 297–301 (2017). [PubMed: 28099430]
16. Mimitou EP et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat. Methods 16, 409–412 (2019). [PubMed: 31011186]
17. Marguerat S et al. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell 151, 671–683 (2012). [PubMed: 23101633]
18. Zheng GXY et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun 8, 14049 (2017). [PubMed: 28091601]
19. O'Huallachain M et al. Ultra-high throughput single-cell analysis of proteins and RNAs by split-pool synthesis. Commun Biol 3, 213 (2020). [PubMed: 32382044]
20. Cao J et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science vol. 357 661–667 (2017). [PubMed: 28818938]
21. Cao J et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361, 1380–1385 (2018). [PubMed: 30166440]
22. Cao J et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502 (2019). [PubMed: 30787437]
23. Rosenberg AB et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 360, 176–182 (2018). [PubMed: 29545511]
24. Lareau CA et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology vol. 37 916–924 (2019).
25. Datlinger P, Rendeiro AF, Boenke T & Krausgruber T Ultra-high throughput single-cell RNA sequencing by combinatorial fluidic indexing. bioRxiv (2019).
26. Heaton H et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat. Methods 17, 615–620 (2020). [PubMed: 32366989]
27. Gehring J, Hwee Park J, Chen S, Thomson M & Pachter L Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nat. Biotechnol 38, 35–38 (2020). [PubMed: 31873215]
28. Ferrer-Font L et al. Panel design and optimization for high-dimensional immunophenotyping assays using spectral flow cytometry. Curr. Protoc. Cytom 92, e70 (2020). [PubMed: 32150355]
29. Collin M & Bigley V Human dendritic cell subsets: an update. Immunology 154, 3–20 (2018). [PubMed: 29313948]

METHOD REFERENCES

1. Zunder ER et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nat. Protoc 10, 316–333 (2015). [PubMed: 25612231]
2. Traag VA, Waltman L & van Eck NJ From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep 9, 5233 (2019) [PubMed: 30914743]

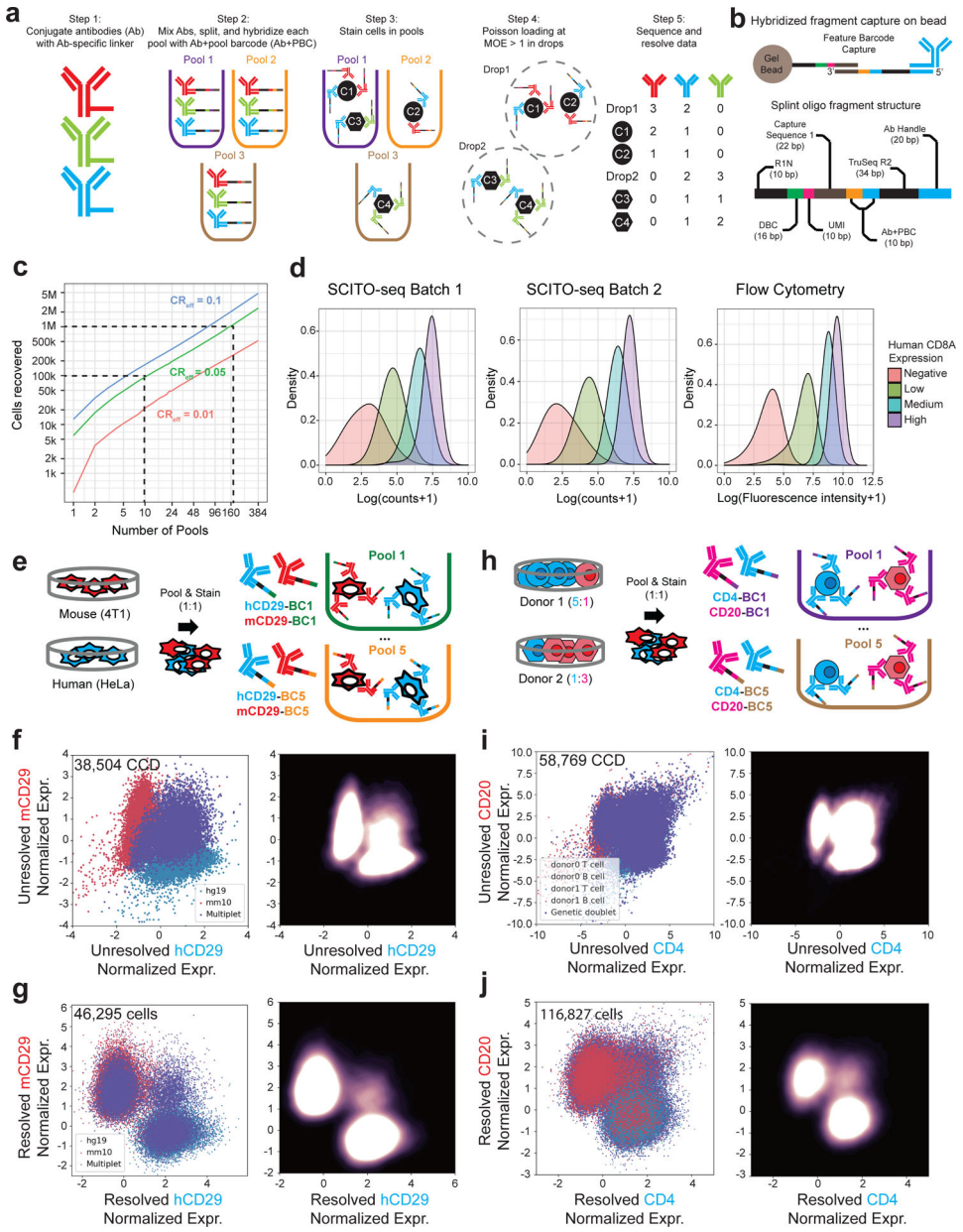


Fig. 1 |. Design of SCITO-seq and mixed-species proof-of-concept.

(a) SCITO-seq workflow. Each antibody is conjugated with a unique antibody barcode (red, green and blue) and hybridized with a splint oligo containing antibody and pool barcodes (Ab+PBC (Pool barcode): [red, blue, green] × [purple, orange, brown]). Cells are split into pools and stained and then mixed and loaded for dsc-seq at high loading concentrations. Cells are resolved from the resulting data using the combinatorial index of Ab+PBC and droplet barcodes. (b) A detailed structure of the SCITO-seq fragment produced. The primary antibody-specific universal oligo is also a hybridization handle. The splint oligo consists of the reverse complement sequence to the handle followed by a TruSeq adaptor (black), the compound Ab+PBC (blue+orange), and a gelbead bound sequence (i.e., the 10× 3’v3 feature barcode capture sequence 1 (brown)). The Ab+PBC and the droplet barcode

(DBC) form a combinatorial index unique to each cell. (c) Cell recovery and collision rate analysis. Number of cells recovered as a function of the number of pools at three commonly accepted collision rates (1%, 5% and 10%). (d) Density histograms of SCITO-seq vs FACS showing 4 different bins of CD8A expression. Log1p transformed SCITO-seq counts for two pools are compared with the log1p fluorescence intensity per cell from FlowJo (v10). (e) Mixed species (HeLa and 4T1) proof-of-concept experiment. HeLa and 4T1 cells are mixed and stained in five separate pools at a ratio of 1:1 with pool-barcoded human and mouse anti-CD29 antibodies. Scatter (left) and density (right) plots of (f) 38,504 unresolved cell-containing droplets (CCD) and (g) 46,295 resolved cells while loading 10^5 cells. (h) Schematic of human mixing experiment where different ratios of T and B cells (5:1 and 1:3) were mixed prior to splitting and staining with five pools of CD4 and CD20 antibodies. Cell types are indicated by color (T: blue, B: red) while shapes indicate donors. Side-by-side scatter plot and density plots of (i) unresolved and (j) resolved cells for loading 2×10^5 cells. Merged ADT counts are generated by summing all counts for each antibody across pools. Resolved data obtained after assigning cells based on Ab+PBC and DBC barcodes.

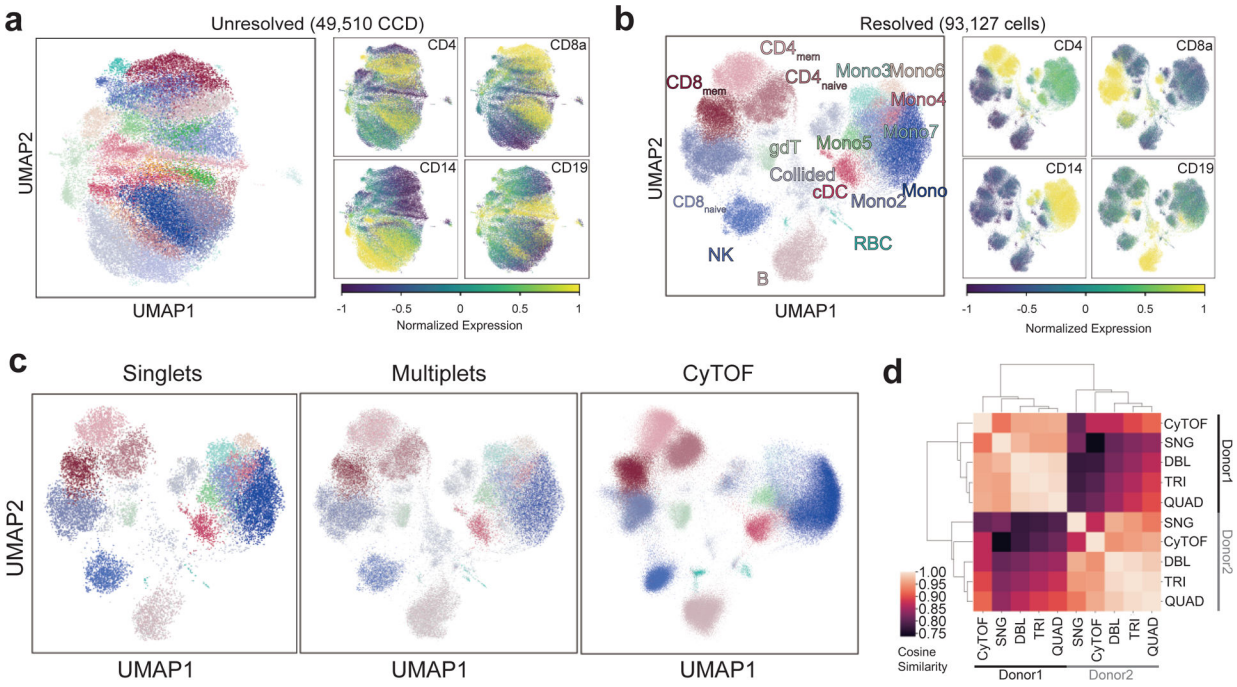


Fig. 2 |. Ultra high-throughput PBMC profiling of healthy controls using SCITO-seq.

(a) UMAP projection of 49,510 CCDs using merged ADT counts of a 28-plex antibody panel showing key lineage markers. (b) UMAP projection of 93,127 PBMCs resolved using Ab+PBCs show the canonical myeloid and lymphoid cell types defined by known markers. (c) UMAP projections of PBMCs resolved from singlets (left), resolved from multiplets (middle), and profiled using CyTOF (right). Principle Component Analysis (PCA)-based integration of data (Ingest function from Scanpy) was used to determine overlapping cell populations between SCITO-seq and CyTOF. (d) Heatmap of pairwise cosine similarity (scaled colors) between estimated cell type proportions for cells originating from singlets (SNG), doublets (DBL), triplets (TRI), quadruplets (QUAD), and CyTOF per donor.

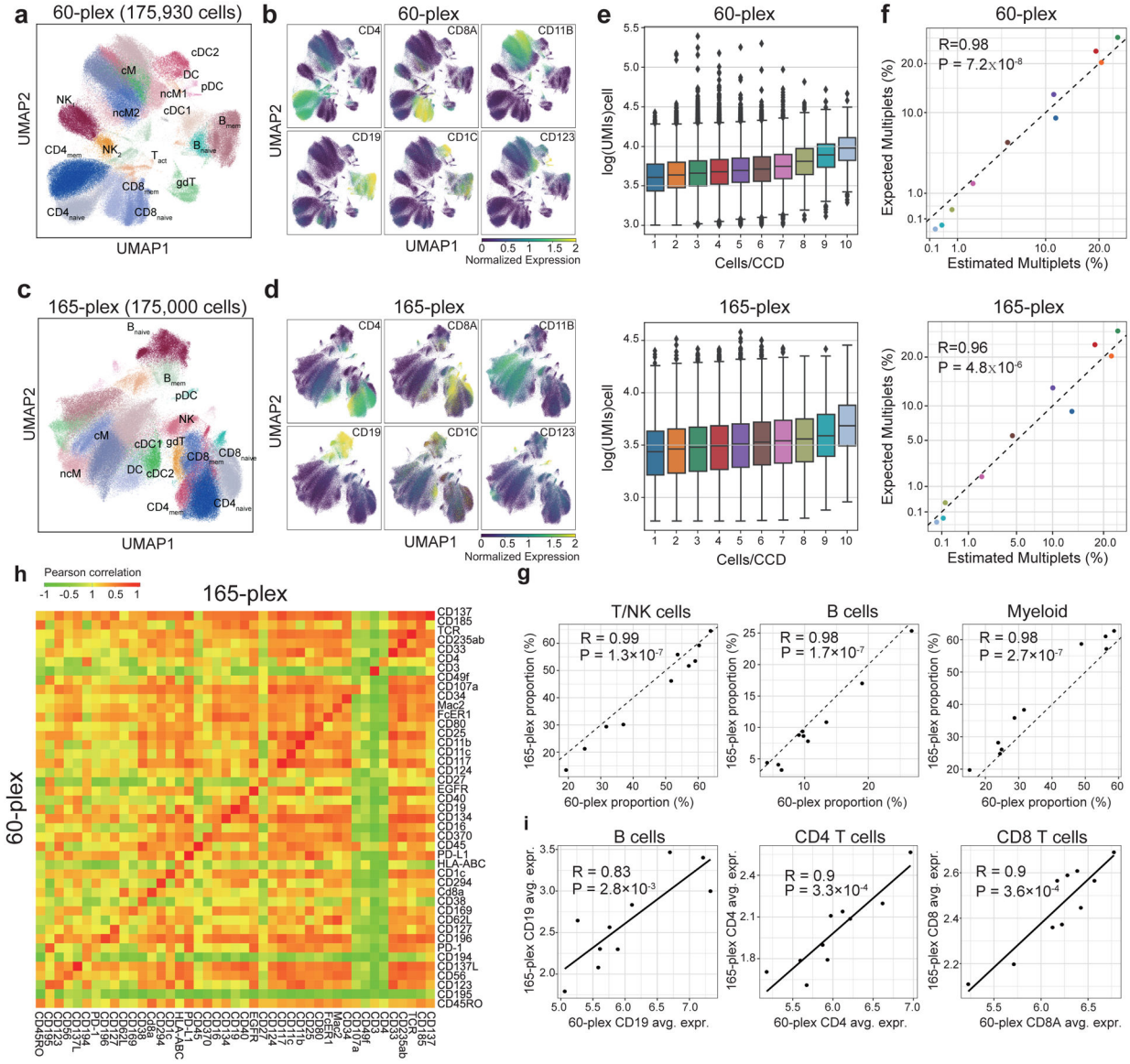


Fig. 3 |. Extending SCITO-seq for compatibility with 60-plex custom and 165-plex commerical antibody panels.

(a) UMAP projection of 175,930 resolved PBMCs using a panel of 60-plex antibodies colored by Leiden clusters and (b) key lineage markers. Subscripts/prefixes stands for: c:conventional, nc:non-conventional, act:activated, gd:gamma-delta. (c) UMAP projection of 175,000 resolved PBMCs using a panel of 165-plex TotalSeq-C antibodies (TSC 165-plex) colored by Leiden clusters and (d) key lineage markers. (e) Distributions of UMIs per cell (y-axis) for CCDs with different numbers of cells (1–10) encapsulated for 60-plex (up, n=6,831, 24,031, 42,274, 44,251, 31,805, 16,094, 6,600, 2,437, 1,068, 538) and TSC 165-plex (down, n=7,779, 25,573, 40,949, 42,036, 28,954, 15,411, 6,957, 3,178, 1,859, 2,304) experiments. Lines are medians, box extends from 25% to 75%, dots are outliers beyond 1.5× interquartile range. (f) Correlation plots for 60-plex (up) and TSC 165-plex (down) experiments comparing estimated (x-axis) and expected multiplet rates (y-axis). Ten

points are shown from 1 to 10 cells encapsulated per CCD and colors are matched to panel (e). (g) Correlations of the cell composition estimates using the 60-plex (x-axis) versus TSC 165-plex (y-axis) experiments for major cell lineages (T and NK cell (left), B cell (middle), myeloid cells (right)) across the same 10 donors represented in each pooled experiment. (h) Row-clustered heatmap of pairwise correlation of 43 overlapping markers between the 60-plex (row) and 165-plex (column) experiments. (i) Correlations of representative marker expressions within specific cell types between the 60-plex (x-axis) and 165-plex (y-axis) experiments (CD19 in B cell, CD4 in CD4 T cells, CD8A in CD8 T cells). The p-values for (f), (g) and (i) is calculated as the corresponding two-sided p-value for the t-distribution with n-2 degrees of freedom. The same 10 donors in 10 pools (1 donor stained in each well) were used for both 60- and 165-plex data.

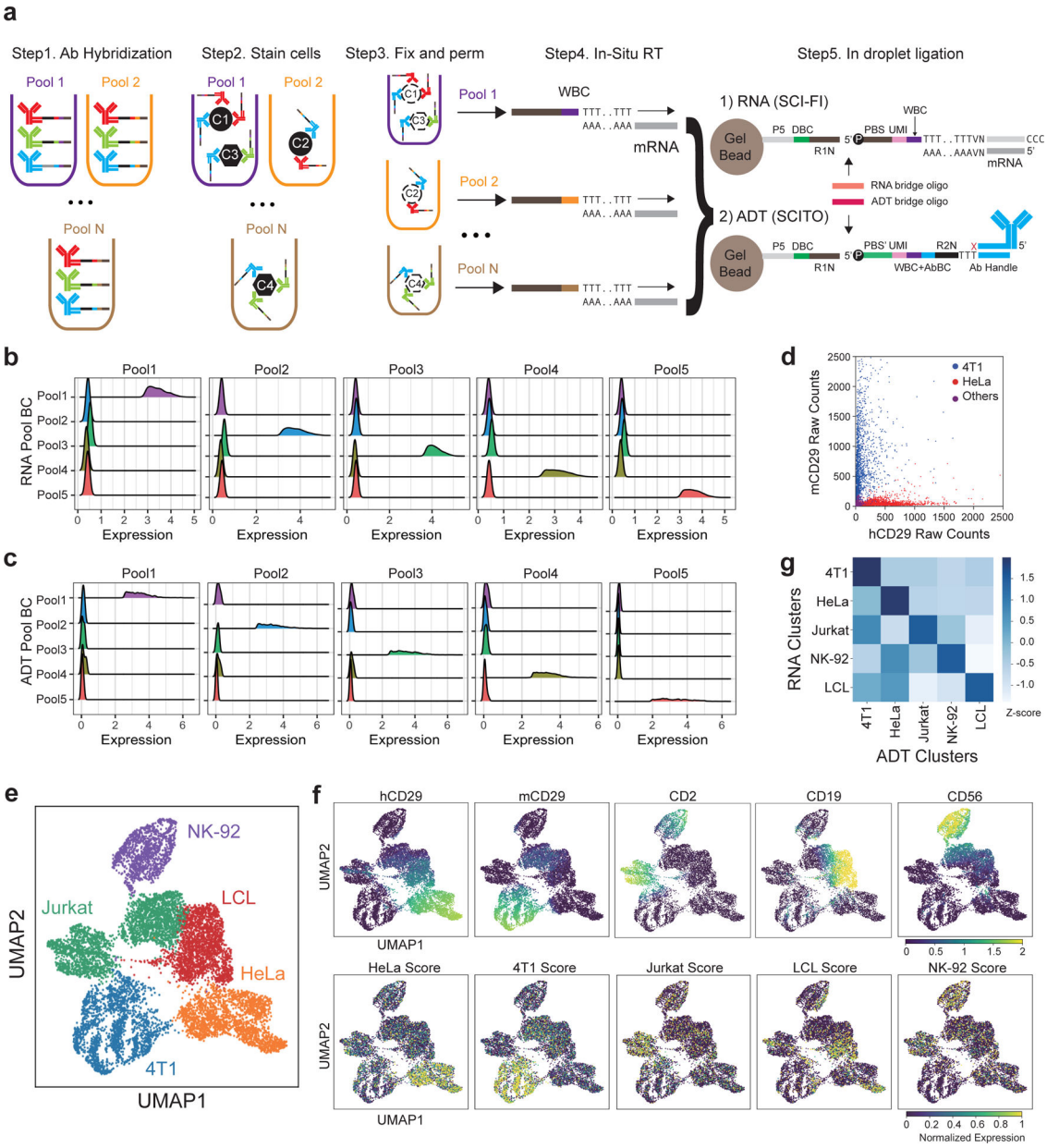


Fig. 4 |. Integrating SCITO-seq and scifi-RNA-seq for simultaneous profiling of transcripts and surface proteins.

(a) Schematic of the SCITO-seq and scifi-RNA-seq coassay. Hybridized SCITO-seq antibodies are used to stain cells in different pools. Cells are washed with buffer then fixed and permeabilized with methanol. Transcripts undergo in-situ reverse transcription (RT) with pool-barcoded RT primers (well barcode denoted as WBC). cDNA and ADT molecules are then captured with RNA- and ADT-specific bridge oligos (orange and red hybridized to PBS and PBS') and ligated to DBCs in droplet (See Extended Data Fig. 9b for details). Ridgeplots of distribution of cells with specific pool barcodes for the (b) RNA library and (c) ADT library. (d) Barnyard plot showing expected staining of human anti-CD29 (x-axis) and mouse anti-CD29 (y-axis) antibodies on HeLa cells and 4T1 cells respectively. Other cell lines are negative for both antibodies as expected. (e) UMAP projection generated from

ADT data colored by Leiden clusters. (f) UMAP projection colored by ADT markers (top) and corresponding cell-line-specific transcriptomic signatures scored using the Scanpy's score genes function (bottom). (g) Heatmap of the overlap analysis of mRNA (y-axis) and ADT markers (x axis), mRNA marker genes are mapped onto cell-type specific ADT clusters (Gene scores are calculated on cells corresponding to ADT based clusters) for all 5 cell lines. The color-scaled values are standardized z–score scale.