Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput

Todd M Gierahn[1,8], Marc H Wadsworth II[2–4,8], Travis K Hughes[2–4,8], Bryan D Bryson[4,5], Andrew Butler[6,7], Rahul Satija[6,7], Sarah Fortune[4,5], J Christopher Love[1,3,4,9] & Alex K Shalek[2,3,4,9]

Single-cell RNA-seq can precisely resolve cellular states, but applying this method to low-input samples is challenging. Here, we present Seq-Well, a portable, low-cost platform for massively parallel single-cell RNA-seq. Barcoded mRNA capture beads and single cells are sealed in an array of subnanoliter wells using a semipermeable membrane, enabling efficient cell lysis and transcript capture. We use Seq-Well to profile thousands of primary human macrophages exposed to *Mycobacterium tuberculosis*.

The emergence of single-cell genomics has enabled new strategies for identifying the cellular and molecular drivers of biological phenomena[1–19]. Patterns in genome-wide mRNA expression measured by single-cell RNA-seq (scRNA-seq) can be leveraged to uncover distinct cell types, states and circuits within cell populations and tissues[1–5,9–13]. To inform our understanding of healthy and diseased behaviors and eventually guide precision diagnostics and therapeutics, we need broadly applicable scRNA-seq methods that are easy to use and enable high-throughput studies of cellular phenotypes, particularly for low-input ($\leq 10^4$ cells) samples such as clinical specimens.

Typically, scRNA-seq involves isolating and lysing individual cells, then independently reverse transcribing and amplifying their mRNAs before generating barcoded libraries that are pooled for sequencing. Although manual picking[2,5,8], FACS sorting[1,3,4] or integrated microfluidic circuits[7,9,10] can isolate single cells, these approaches are constrained in scale by cost, time and labor. Recently developed massively parallel methods assign unique barcodes to each cell's mRNAs during reverse transcription, enabling ensemble processing while retaining single-cell resolution. These techniques typically yield single-cell libraries of lower complexity, but higher throughput reduces the impact of the technical and intrinsic noise associated with each cell in analyses[11,12]. The most commonly used approach relies on microfluidic devices to generate reverse-emulsion droplets that couple single cells with uniquely barcoded mRNA capture beads[11,12]. Droplet-based techniques, however, can have inefficiencies in encapsulation, introduce technical noise through differences in cell lysis time and require specialized equipment—limiting where, when and at what scale scRNA-seq can be performed.

One alternative to droplets is to use arrays of subnanoliter wells loaded by gravity, which reduces the need for peripheral equipment, decreases dead volumes and facilitates parallelization. As a proof of principle, cells and beads have been coconfined in nanowell arrays to perform targeted single-cell transcriptional profiling[13], yet the absence of a seal significantly impairs capture efficiency and increases cross-contamination (**Supplementary Fig. 1**). Nanowells have also been combined with microfluidic channels that facilitate oil-based single-cell isolation via fluid exchange[14]. Nevertheless, this design limits buffer exchange and necessitates integrated temperature and pressure controllers, impacting ease of use and portability[15]. Semiporous-membrane-covered nanowells have been used to link pairs of specific transcripts from single cells[16]; however, this approach used many beads per well, precluding the creation of unique single-cell libraries, and transcript capture and sealing efficiency were not addressed.

To overcome these challenges, we developed Seq-Well, a simple, portable platform for massively parallel scRNA-seq (**Supplementary Fig. 2**). Seq-Well confines single cells and barcoded poly(dT) mRNA capture beads in a PDMS array of ~86,000 subnanoliter wells. Wells accommodate only one bead, enabling single-bead loading efficiencies of ~95% (**Fig. 1a**, **Supplementary Fig. 3a** and **Supplementary Video 1**). A simplified cell-loading scheme in turn permits capture efficiencies of around 80% (see Online Methods and **Supplementary Fig. 3b**), with a dual occupancy rate that can be tuned by adjusting the number of cells loaded and visualized before processing (**Supplementary Fig. 3c**).

A key unique feature of Seq-Well is the use of selective chemical functionalization to facilitate reversible attachment of a semipermeable polycarbonate membrane (10-nm pore size) in physiologic buffers. This enables rapid solution exchange for efficient cell lysis but traps biological macromolecules for improved transcript capture and reduced cross-contamination (**Supplementary Fig. 4a**, **Supplementary Protocol** and **Supplementary Video 2**). The array's three-layer surface functionalization comprises an aminosilane base[20] crosslinked to a bifunctional poly(glutamate)–chitosan top via a p-phenylene diisothiocyante intermediate (see

[1]Koch Institute for Integrative Cancer Research, MIT, Cambridge, Massachusetts, USA. [2]Institute for Medical Engineering & Science (IMES) and Department of Chemistry, MIT, Cambridge, Massachusetts, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA. [5]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA. [6]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, USA. [7]New York Genome Center, New York, New York, USA. [8]These authors contributed equally to this work. [9]These senior authors contributed equally to this work. Correspondence should be addressed to A.K.S. (shalek@mit.edu) or J.C.L. (clove@mit.edu).

RECEIVED 25 JULY 2016; ACCEPTED 28 DECEMBER 2016; PUBLISHED ONLINE 13 FEBRUARY 2017; DOI:10.1038/NMETH.4179

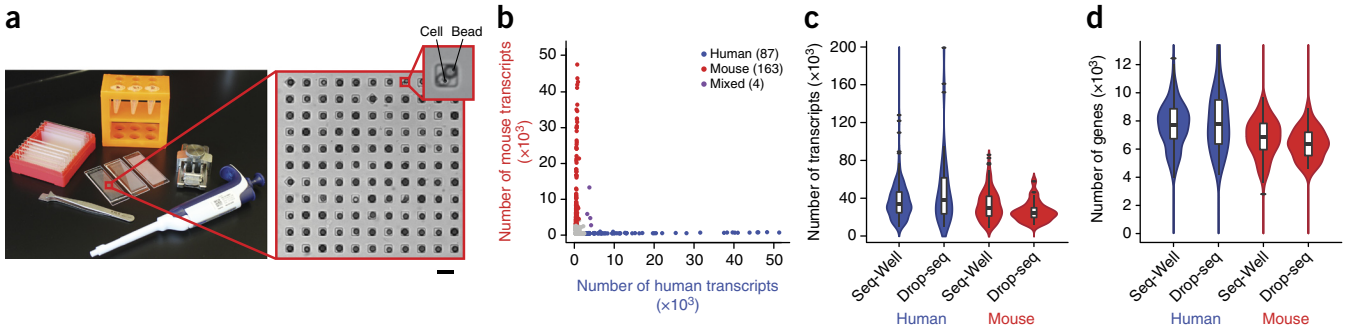

Figure 1 | Seq-Well: a portable, low-cost platform for high-throughput single-cell RNA-seq of low-input samples. (**a**) Equipment and array used to capture and lyse cells, respectively. Scale bar, 100 µm. (**b**) Sequencing a mix of human (HEK293) and mouse (NIH/3T3) cells reveals distinct transcript mapping and single-cell resolution. Cells with >2,000 human and <1,000 mouse transcripts are labeled as human, and cells with >2,000 mouse and <1,000 human transcripts are labeled as mouse. Of the 254 cells identified, 4 (1.6%) had a mixed phenotype. (**c**,**d**) Number of transcripts (**c**) and genes (**d**) detected in single-cell libraries generated by Seq-Well or Drop-seq (ref. 12; center-line: median; limits, first and third quartile; whiskers, ±1.5 IQR; points; values, >1.5 IQR). Using Seq-Well (Drop-seq), an average of 37,878 (48,543) transcripts or 6,927 (7,175) genes were detected among human HEK cells (n = 159 for Seq-Well; n = 48 for Drop-seq); and an average of 33,586 (26,700) transcripts or 6,113 (5,753) genes were detected among mouse 3T3 cells (n = 172 for Seq-Well; n = 27 for Drop-seq) at an average depth of 164,238 (797,915) reads per HEK cell and 152,488 (345,117) reads per 3T3 cell.

Online Methods and **Supplementary Fig. 4**). In the outer layer, poly(glutamate) at the inner nanowell surfaces prevents nonspecific binding of mRNAs, while chitosan on the array's top surface encourages efficient sealing to the membrane (see Online Methods; **Supplementary Protocol** and **Supplementary Fig. 4b**,**c**). To test sealing and buffer exchange, we monitored the fluorescence of dye-labeled, cell-bound antibodies before and after adding a guanidinium-based lysis buffer. We observed rapid diffusion of the antibodies throughout the wells within 5 minutes of buffer addition and—unlike in unsealed or previously described, membrane-covered BSA-blocked arrays[16]—we observed little change in fluorescent signal over 30 min, suggesting robust retention of biological macromolecules despite the use of a strong chaotrope (see Online Methods and **Supplementary Fig. 5**).

After lysis, cellular mRNAs are captured by bead-bound poly(dT) oligonucleotides that also contain a universal primer sequence, a cell barcode and a unique molecular identifier (UMI) (see Online Methods and **Supplementary Table 1**). Next, the membrane is peeled off, and the beads are removed for subsequent bulk reverse transcription, amplification, library preparation and paired-end sequencing, as previously described[12] (see Online Methods). Critically, beyond a disposable array and membrane, Seq-Well only requires a pipette, a manual clamp, an oven and a tube rotator to achieve stable, barcoded single-cell cDNAs (**Fig. 1a**), so it can be performed almost anywhere.

To assess transcript capture efficiency and single-cell resolution, we profiled a mixture of 5×10^3 human (HEK293) and 5×10^3 mouse (3T3) cells using Seq-Well. The average fraction of reads mapping to exonic regions was 77.5% (**Supplementary Fig. 6**), demonstrating high-quality libraries. Shallow sequencing from a fraction of an array revealed highly organism-specific libraries, suggesting single-cell resolution and minimal cross-contamination (**Fig. 1b** and **Supplementary Fig. 7a–c**). In the absence of membrane sealing, by comparison, we obtained poor transcript and gene detection as well as substantial cross-contamination (**Supplementary Fig. 1**). From deeper sequencing of a fraction of a second array, we detected an average of 37,878 mRNA transcripts from 6,927 genes in HEK cells and 33,586 mRNA transcripts from 6,113 genes in 3T3 cells, comparable to a droplet-based approach using the same mRNA capture beads (Drop-seq)[12] (**Fig. 1c**,**d** and **Supplementary Figs. 7** and **8**). Upon downsampling to match read depths, we also observed levels of transcript and gene detection consistent with those of other massively parallel bead-based scRNA-seq methods (see Online Methods and **Supplementary Fig. 7d–g**). Moreover, bulk RNA-seq data were strongly correlated with populations constructed *in silico* from individual HEK cells ($R = 0.751 \pm 0.073$ to $R = 0.983 \pm 0.0001$ for populations of 1–1,000 single cells, respectively), suggesting representative cell and transcript sampling (see Online Methods and **Supplementary Fig. 9**).

To examine Seq-Well's ability to resolve populations of cells in complex primary samples, we loaded human peripheral blood mononuclear cells (PBMCs) into arrays in triplicate before beads, allowing us to perform on-array multicolor imaging cytometry (see Online Methods; **Fig. 2a**,**b**; and **Supplementary Tables 2** and **3**). Sequencing one-third of the beads recovered from each array yielded 3,694 high-quality single-cell libraries (see Online Methods). Unsupervised graph-based clustering revealed unique subpopulations corresponding to major PBMC types (see Online Methods, **Fig. 2b**, **Supplementary Figs. 10–12** and **Supplementary Table 4**). Each array yielded similar subpopulation frequencies (**Fig. 2c**), with detection efficiencies comparable to those of other massively parallel technologies (**Supplementary Fig. 13**). The proportion of each subpopulation determined by sequencing also matched on-array immunophenotyping results (**Fig. 2a**,**b**). Critically, sequencing provided additional information; in addition to resolving dendritic cells from monocytes (**Fig. 2b**), we found significant variation among the monocytes (captured in PC3) due to differential expression of inflammatory and antiviral gene programs (**Fig. 2d**)[1,3]. Our results show that characterizing a sample in two ways using a single platform increases the amount of information that can be extracted from a precious specimen, while allowing analysis of one measurement to be interpreted in the context of the other.

Finally, to test the portability of Seq-Well, we profiled primary human macrophages exposed to *Mycobacterium tuberculosis*

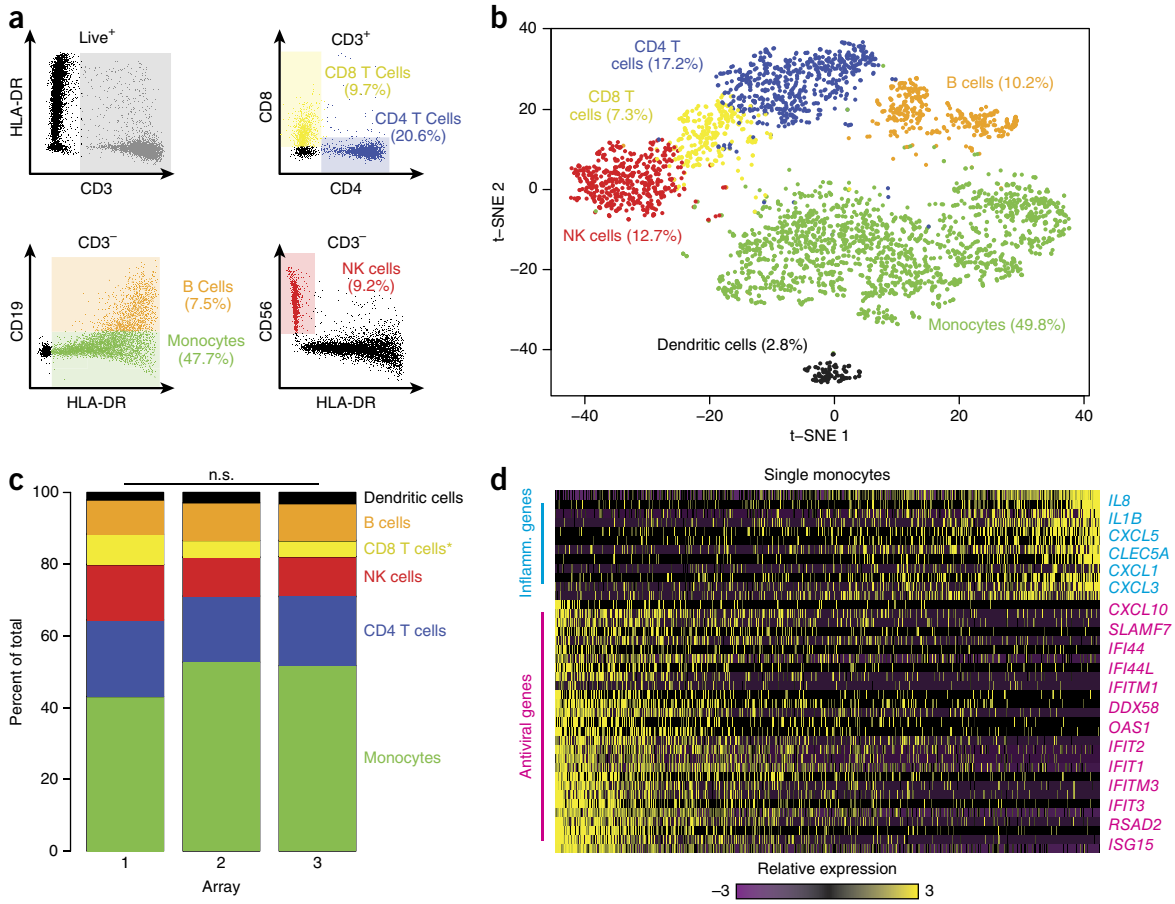

Figure 2 | Combined image cytometry and scRNA-seq of human PBMCs. (**a**) Hierarchical gating scheme used to analyze PBMCs labeled with a panel of fluorescent antibodies, loaded onto three replicate arrays and imaged before bead loading (see Online Methods). Myeloid cells (green) were identified as the population of hCD3(–) HLA-DR(+) CD19(–) cells; B cells (orange) as the subset of hCD3(–) HLA-DR(+) CD19 (+) cells; CD4 T cells (blue) as the subset of CD3(+) CD4(+) cells; CD8 T cells (yellow) as the CD3(+) CD8(+) subset of cells; and NK cells (red) as the subset of CD3(–) HLA-DR(–) CD56(+) CD16(+) cells. (**b**) t-SNE visualization of clusters identified among 3,694 human Seq-Well PBMCs single-cell transcriptomes recovered from the imaged array and the two additional arrays (see Online Methods and **Supplementary Figs. 10–12**). Clusters (subpopulations) are labeled based on annotated marker gene (**Supplementary Fig. 10**). (**c**) Distribution of transcriptomes captured on each of the biological replicate arrays, run on separate fractions of the same set of PBMCs. No shifts are statistically significant (n.s. = not significant; see Online Methods) except for a slightly elevated fraction of CD8 T cells in array 1 (*, $P = 1.0 \times 10^{-11}$; Chi-square test, Bonferroni corrected). (**d**) Relative expression level of a set of inflammatory and antiviral genes among cells identified as monocytes. Inflamm., inflammatory.

(H37Rv) in a BSL3 facility (see Online Methods). In total, we recovered 14,218 macrophages (from a total of 40,000 loaded across experiments) with greater than 1,000 mapped transcripts from an *M. tuberculosis*-exposed and an unexposed array. Unsupervised analysis of 4,638 cells with greater than 5,000 transcripts per cell revealed five distinct clusters (**Fig. 3a**,**b**; **Supplementary Fig. 14a**,**b** and **Supplementary Table 5**). Two clusters had lower transcript capture and high mitochondrial gene expression (suggestive of low-quality libraries)[17] and were removed; the remaining three (2,560 cells) were identified in both the exposed and unexposed samples (**Fig. 3a**, **Supplementary Figs. 14c**,**d** and **15**), and they likely represent distinct subphenotypes present in the initial culture.

We next examined common and cluster-specific gene enrichments (see Online Methods). Although clusters 1 and 3 did not present strong stimulation-independent enrichments, cluster 2 uniquely expressed several genes associated with metabolism (**Supplementary Tables 6** and **7**). Intriguingly, within each cluster we observed pronounced shifts in gene expression in response to *M. tuberculosis* (see Online Methods, **Fig. 3c** and **Supplementary Table 8**), with common enrichments for gene sets previously observed in response to intracellular infection, LPS stimulation and activation of TLR7/8 receptor (**Supplementary Tables 9** and **10**). Cluster 1 uniquely displayed stimulation-induced shifts in several genes associated with cell growth, cluster 3 in transcripts associated with hypoxia, and cluster 2 (again) in genes linked to metabolism. Overall, these data suggest that basal cellular heterogeneity may influence ensemble *M. tuberculosis* responses. Equally important, they demonstrate Seq-Well's ability to acquire large numbers of single-cell transcriptomes in challenging experimental environments.

In conclusion, Seq-Well is a robust platform for scalable, single-cell transcriptomics applicable to almost any cellular suspension for which a reference genome or transcriptome exists. The technique is inexpensive, user friendly, portable, and efficient; it enables scRNA-seq to accelerate scientific and clinical discovery even when working with limited samples. Furthermore, the ability to measure protein secretion and cell-surface

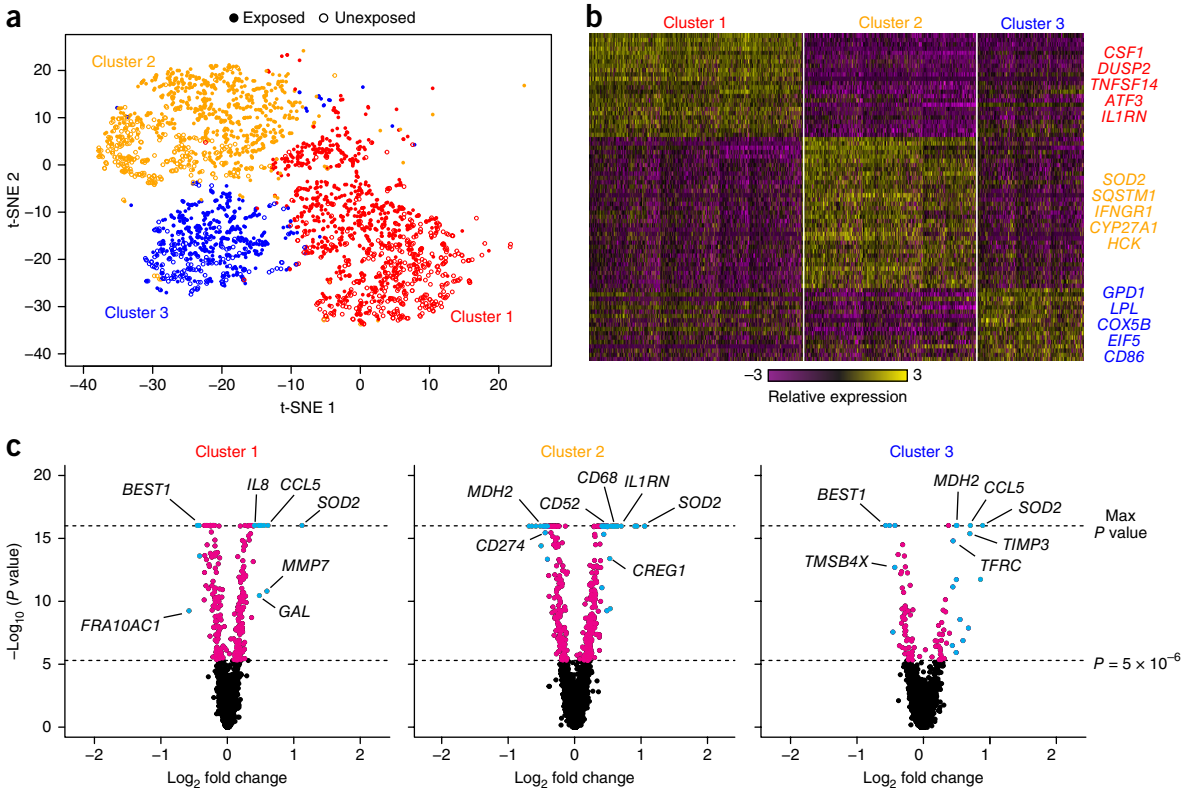

Figure 3 | Sequencing of TB-exposed macrophages in a BSL3 facility using Seq-Well. (**a**) t-SNE visualization of single-cell clusters identified among 2,560 macrophages (1,686 exposed, solid circles; 874 unexposed, open circles) generated using five principal components across 377 variable genes (see Online Methods). (**b**) Marker genes for the three phenotypic clusters of macrophages highlighted in **a**. (**c**) Differential expression between exposed and unexposed macrophages within each cluster showing genes enriched in cells exposed to *M. tuberculosis*. Cyan, genes with *P* values less than 5.0×10^{-6} (threshold for statistical significance, determined by a likelihood ratio test) and absolute $\log_2$ fold changes greater than 0.4 (threshold used for differential expression). Magenta, genes with *P* values less than 5.0×10^{-6} but absolute $\log_2$ fold changes less than 0.4. Black, remaining genes.

expression on the same platform[18,19] foreshadows multi-omic single-cell measurements at scale.

METHODS

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

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

ACKNOWLEDGMENTS

We thank K. Shekhar, T. Tickle, and M. Xie for fruitful discussions. This work was supported by the Searle Scholars Program (A.K.S.), the Beckman Young Investigator Program (NIH, A.K.S.), NIH New Innovator Award DP2 OD020839 (A.K.S.), NIH U24 AI11862-01 (A.K.S.), P50 HG006193 (A.K.S.), the Bill and Melinda Gates Foundation grant 03629000189 (A.K.S., J.C.L. and S.F.), the Ragon Institute (A.K.S. and S.F.), the Burroughs Wellcome Foundation (S.F.), NIH P30 AI060354 (S.F.), DP3 DK09768101 (NIH, J.C.L.), P01 AI045757 (J.C.L.), NIH R21 AI106025 (J.C.L.), NIH R56 AI104274 (J.C.L.), the W.M. Keck Foundation (J.C.L.), and the US Army Research Office through the Institute for Soldier Nanotechnologies, under contract number W911NF-13-D-0001 (J.C.L.). This work was also supported in part by the Koch Institute Support (core) NIH Grant P30-CA14051 from the National Cancer Institute. J.C.L. is a Camille Dreyfus Teacher–Scholar.

AUTHOR CONTRIBUTIONS

T.M.G., M.H.W. II, T.K.H., J.C.L. and A.K.S. developed the concepts and designed the study. T.M.G., M.H.W. II, T.K.H. and B.D.B. performed the experiments. T.M.G., M.H.W., T.K.H., A.B. and R.S. performed bioinformatic analysis. S.F. helped design and interpret the *M. tuberculosis* experiments. All authors analyzed and interpreted the data. T.M.G., M.H.W. II, T.K.H., J.C.L. and A.K.S. wrote the manuscript with feedback from all authors.

COMPETING FINANCIAL INTERESTS

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

1. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).
2. Lohr, J.G. *et al. Nat. Biotechnol.* **32**, 479–484 (2014).
3. Shalek, A.K. *et al. Nature* **510**, 363–369 (2014).
4. Tirosh, I. *et al. Science* **352**, 189–196 (2016).
5. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
6. Bendall, S.C. *et al. Science* **332**, 687–696 (2011).
7. Buenrostro, J.D. *et al. Nature* **523**, 486–490 (2015).
8. Smallwood, S.A. *et al. Nat. Methods* **11**, 817–820 (2014).
9. Zeisel, A. *et al. Science* **347**, 1138–1142 (2015).
10. Treutlein, B. *et al. Nature* **509**, 371–375 (2014).
11. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
12. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
13. Fan, H.C., Fu, G.K. & Fodor, S.P. *Science* **347**, 1258367 (2015).
14. Bose, S. *et al. Genome Biol.* **16**, 120 (2015).
15. Yuan, J. & Sims, P.A. *Sci. Rep.* **6**, 33883 (2016).
16. DeKosky, B.J. *et al. Nat. Biotechnol.* **31**, 166–169 (2013).
17. Ilicic, T. *et al. Genome Biol.* **17**, 29 (2016).
18. Yamanaka, Y.J. *et al. Integr. Biol. (Camb.)* **4**, 1175–1184 (2012).
19. Han, Q. *et al. Proc. Natl. Acad. Sci. USA* **109**, 1607–1612 (2012).
20. Steinberg, G., Stromsborg, K., Thomas, L., Barker, D. & Zhao, C. *Biopolymers* **73**, 597–605 (2004).

ONLINE METHODS

Running Seq-Well. See *Protocol Exchange*[21], **Supplementary Protocol** and **Supplementary Videos 1** and **2** or http://www.shaleklab.com for a step-by-step protocol for Seq-Well. For this work, we used the primers listed in **Supplementary Table 1**.

Bead synthesis. Barcoded oligo-dT beads (as described in Macosko *et al.*[12]) were purchased from Chemgenes (Wilmington, Massachusetts, USA; cat. no. MACOSKO-2011-10) at 10 umol scale (~100 arrays). Bead functionalization and reverse phosphoramidite synthesis was performed by Chemgenes Corporation using Toyopearl HW-65S resin (30 micron mean particle diameter) obtained from Tosoh Biosciences (cat. no. 19815). Surface hydroxyls were reacted with a PEG derivative to obtain an 18-carbon linker to serve as a support for oligo synthesis. Reverse-direction phosphoramidite synthesis was performed using an Expedite 8909 DNA/RNA synthesizer at 10 micromole scale with a coupling time of 3 min. Initially, a conserved PCR handle was synthesized followed by 12 rounds of split and pool synthesis to generate 16,777,216 unique barcode sequences. Addition of an 8-mer random sequence was performed to generate unique molecular identifiers (UMIs) on each capture oligo. Finally, a 30-mer poly-dT capture sequence was synthesized to enable capture of polyadenylated mRNA species.

Imaging differential surface functionalization. Differential labeling of the top and inner well surfaces was visualized by substituting 1 ug/mL PE–strepavidin for chitosan (step 8, Seq-Well Protocol[21]) and 1 ug/mL AlexaFluor488-Streptavidin for the polyglutamate (step 10, Seq-Well Protocol[21]) in the standard functionalization protocol (**Supplementary Fig. 3**). Carboxylation of the inner well surfaces was visualized by treating the functionalized array with 100 μg/mL EDC/10 μg/mL NHS MES (pH 6.0) solution for 10 min, washing twice with MES buffer, once with sodium borate buffer (pH 8.5), and incubating overnight with 1 μg/mL Alexa-Fluor 568-labeled antibody. Arrays were washed three times with phosphate-buffered saline (PBS) and imaged using Alexa Fluor 568 channel (**Supplementary Table 2**).

Visualizing lysate retention (imaging). PBMCs were labeled with αCD45–AF647 (BioLegend 304020, diluted 1:20). Cells were washed and loaded onto two arrays previously blocked with 1% BSA solution for 30 min and one array functionalized with chitosan as described above. A polycarbonate membrane was attached to the chitosan-functionalized array as described above. The array was submerged in PBS and imaged for AF647 fluorescence to identify wells containing cells. The BSA-blocked arrays were imaged before membrane attachment because the membrane would detach when submerged in media. After imaging, a plasma-treated polycarbonate membrane was attached to one of the BSA-blocked arrays as described[16]. Briefly, the membrane was placed on the array with forceps, and all excess media were aspirated from the array. The open BSA-blocked array and the chitosan array were submerged in 5 mL of 5 M GCTN lysis buffer. 500 μL of lysis buffer was placed on the top of membrane attached to the BSA-blocked array as described[16]. 5 and 30 minutes later, 100 block positions were imaged on each array, encompassing 12,100 individual wells. Automated image analysis software was used to background subtract each image, identify cell and well locations and extract AF647 signal intensity of the cells and the well volumes (**Supplementary Fig. 4**).

Calculating bead loading efficiency. Bead loading efficiencies were determined by loading two functionalized arrays with beads as outlined above (**Supplementary Fig. 1**). Arrays were imaged in transmitted light and AF488 channel (**Supplementary Table 2**) to capture bead autofluorescence. Automated image analysis was used to identify well locations and extract the 75th percentile fluorescence intensity in each well. Histogram analysis of fluorescence intensities was used to identify empty wells and wells containing beads. Finally, manual review of 50 randomly selected image positions, each containing 121 nanowells, of a total of 690 positions was used to calculate the frequency of wells containing two beads.

Calculating cell loading efficiency. To calculate cell loading efficiencies and well occupancy distributions (**Supplementary Fig. 3**), HEK293 and 3T3 cells were labeled with Calcein AM (Life Technologies) and Calcein Violet (Life Technologies), respectively, per the manufacturer's recommendations. 200 μL of serial dilutions of a 1:1 mix of the cells at an estimated concentration of 1,000, 10,000 and 100,000 cells/mL were loaded in functionalized arrays in triplicate using the standard protocol. To determine the distribution of cells present in 200 μL of these solutions, the same volume of each solution was added to 12 wells of a 96-well plate. 690 array positions on each array were imaged in the transmitted light, AF488 and AF405 spectral channels (**Supplementary Table 2**). Overlapping images of each well of the 96-well plate were acquired in the same channels. Automated image analysis was used to identify well and cell locations in the array images. The overlapping images of the 96-well plate were stitched together based on x–y location of each image and analyzed in a similar manner to identify cell locations. All three dilutions were used to determine the distribution of well occupancy as a function of the number of cells loaded. The 10,000 cells/mL dilution were used to calculate cell loading efficiency.

Species-mixing experiments. Murine NIH/3T3 cells (ATCC, CRL-1658) were cultured in Dulbelcco's modified Eagle's medium (DMEM) with glutamate and supplemented with 10% fetal bovine serum (FBS) at 37 °C and 5% CO_2. Human 293T cells (ATCC, CRL-11268) were cultured at 37 °C and 5% CO_2 in DMEM with glutamate supplemented with 10% FBS. The media were removed from the culture flasks, which were then rinsed with 5 mL of 1× PBS. Cells were detached from the surface of the culture flasks by applying 3.5 mL of Trypsin-LE (Life Technologies) and incubating at room temperature for 5 min. Once cells had deadhered, 10 mL of complete media was added, and cells were pelleted by spinning at 500× g for 10 min. Cell pellets were resuspended in 1 mL of media, and a 10 μL aliquot was used to count cells. A total of 100,000 HEK and 3T3 cells were again pelleted and resuspended in 1 mL of media. For species-mixing experiments, a total of 200 μL of a single-cell suspension containing 5,000 HEK and 5,000 NIH/3T3 cells was applied to the surface of two nanowell devices loaded with beads. In the first experiment, of the 60,000 beads collected from the array, 9,600 beads were pooled for subsequent processing and sequencing, from which we identified 254 high-quality cells with greater than 2,000 transcripts. In the second

experiment, of the 25,000 beads collected from the array, 15,000 beads were pooled for subsequent processing and sequencing, from which we identified 331 high-quality cells with greater than 10,000 transcripts, greater than 2,000 genes, and greater than 90% transcript purity (i.e. >90% of transcripts from the same species). Also, as in Drop-Seq, we attempted to validate capture efficiency using ERCC spike-ins; however, this required us to load ERCCs onto the nanowell array by pipetting, which proved inefficient to properly assess capture efficiency since we could not evenly distribute ERCCs to nanowells.

HEK population experiments. HEK293 cells were cultured in RPMI supplemented with 10% FBS. A total of 10,000 HEK293 cells were applied to a Seq-Well device and scRNA-seq libraries were generated from 24,000 beads and sequenced on a NextSeq 500. For the bulk RNA-seq sample, cellular lysate from 40,000 HEK293 cells in 200 µL of lysis buffer (5 M GTCN, 1% 2-mercaptoethanol, 1 mM EDTA, and 0.1% Sarkosyl in 1× PBS, pH 6.0) was combined with 40,000 mRNA capture beads in a PCR tube and rotated end over end for 1 h. Afterward, the beads were washed, and a population sequencing library was constructed in an identical manner to that of the single-cell Seq-Well libraries but with reads from the different bead barcodes combined into one population. *In silico* populations were created by randomly sampling 1, 10, 100 or 1,000 HEK cells from a total of 1,453 cells with greater than 3,000 transcripts obtained from a Seq-Well array. Average Pearson correlation coefficients and their s.d. were calculated between 100 randomly generated *in silico* populations for each number of cells and the bead population (**Supplementary Fig. 9**).

Human PBMC experiments. Leukocytes isolated from a leukocyte reduction filter used during platelet aphoresis were purchased from Key Biologics (Memphis, Tennessee). The cells were shipped overnight at room temperature. PBMCs were isolated from the sample using a Ficoll–Hypaque (GE) gradient, washed two times with HBSS buffer, and frozen in 90% FBS/10%DMSO in aliquots of 10^7 cells. The day before the experiment, an aliquot was thawed and rested overnight in RPMI-1640 supplemented with 10% FBS, Pen–Strep, nonessential amino acids, sodium pyruvate, and HEPES buffer (RP10) at 10^6 cells/mL in a 50 mL conical tube. Cells were counted the next day, and 5×10^5 cells were pelleted, resuspended in 1 mL of CellCover solution, and processed as described above.

Array loading for imaging (PBMCs). To quantify cell surface marker protein expression levels on array (**Fig. 2a**), PBMCs were loaded first and imaged before bead addition to avoid potential detection issues associated with bead autofluorescence. Here, cells were resuspended in cold CellCover (Anacyte), an RNA stabilization reagent, and placed at 4 °C for 1 h. Cells were spun down and resuspended in a cocktail containing αCD45-AF647 (BioLegend; HI30), αCD3-PerCP (BioLegend; UCHT1), αCD4-PECy5.5(eBioscience; SK3), αCD56-PECy5(BD Biosciences; B159), αCD8-APCCy7 (BioLegend; RPA-T8), αHLA-DR-PECy7 (BD Biosciences; L243), and αCD19-PE (BioLegend; HIB19) with all antibodies diluted 1:20 in RP10 media and were incubated at 4 °C for 30 min. Cells were washed twice with PBS and resuspended in CellCover10 buffer (CellCover supplemented with 10% FBS and 100 mM sodium carbonate (pH 10) buffer). Functionalized arrays were washed with 5 mL of CellCover10 buffer. 2.0×10^4 cells were loaded onto the array and washed twice with CellCover10 buffer, and finally the array was placed in 5 mL CellCover. Arrays were imaged with a Zeiss AxioVision microscope with Lumencor light source and EMCCD camera using the settings described in **Supplementary Table 2**. Automated imaging software was used to identify cell locations within the images and extract signal intensities in each spectral channel. To generate spillover coefficients for each fluorophore, α–mouse beads (Bangs Labs) were stained individually with each antibody using the same protocol as the cells. Images of the singly stained beads were used to generate spillover coefficients for each fluorophore that were then used to calculate the amount of each fluorophore on each cell as previously described[22]. After imaging, arrays were washed with 5 mL CellCover10 media. Barcoded beads suspended in CellCover10 media were loaded into the array through gentle agitation. Arrays were washed 3x with CellCover10 without FBS and finally washed with CellCover. Arrays were then moved on to membrane attachment.

Human monocyte isolation. Primary human monocytes were isolated from deidentified human buffy coats obtained from the Massachusetts General Hospital Blood Bank using a standard Ficoll gradient and subsequent CD14 positive selection (Stemcell Technologies). Enriched monocytes were cultured in low-adherence flasks (Corning) for 9 d with RPMI media (Invitrogen) supplemented with 10% heat-inactivated FCS (Sigma-Aldrich).

***Mycobacterium tuberculosis* culture.** *Mycobacterium tuberculosis* (Mtb) H37Rv expressing the E2-Crimson fluorescent protein was grown in Difco Middlebrook 7H9 media supplemented with 10% OADC, 0.2% glycerol, 0.05% Tween-80 and Hygromycin B (50 ug/mL).

Macrophage infection and flow cytometry. The Mtb culture was pelleted by centrifugation and washed once with RPMI + 10% FCS, sonicated briefly, and filtered through a 5 µm syringe filter. Monocyte-derived macrophages (MDM) were infected at an MOI of 10 for 4 h and then washed 3× with RPMI + 10% FCS. 24 h after infection, cells were washed briefly with 1× PBS. 10× Trypsin (Life Technologies) was added, and cells were incubated briefly at 37 °C to allow for cell detachment. Detached cells were spun down and resuspended in 1× PBS supplemented with 2% FCS and 1 mM EDTA and then passed through a mesh filter to eliminate clumps. Uninfected and infected cells were sorted by flow cytometry on an Aria IIu flow cytometer. Mtb-infected cells were identified by the presence of an E2-Crimson signal above the background autofluorescence of uninfected cells.

Transcriptome alignment and barcode collapsing. Read alignment was performed as in Macosko *et al.*[12]. Briefly, for each NextSeq sequencing run, raw sequencing data was converted to FASTQ files using bcl2fastq2 that were demultiplexed by Nextera N700 indices corresponding to individual samples. Reads were first aligned to both HgRC19 and mm10, and individual reads were tagged according to the 12-bp barcode sequence and the 8-bp UMI contained in read 1 of each fragment. Following alignment, reads were binned and collapsed onto 12-bp cell barcodes

doi:10.1038/nmeth.4179

that corresponded to individual beads using Drop-seq tools (http://mccarrolllab.com/dropseq). Barcodes were collapsed with a single-base error tolerance (Hamming distance = 1), with additional provisions for single insertions or deletions. An identical collapsing scheme (Hamming distance = 1) was then applied to UMIs to obtain quantitative counts of individual mRNA molecules. Quality metrics are presented in **Supplementary Figures 5** and **8**.

Data normalization. Digital gene expression matrices were obtained by collapsing filtered and mapped reads for each gene by 8-bp UMI sequences within each cell barcode. For each cell, we performed library-size normalization. UMI-collapsed gene expression values for each cell barcode were scaled by the total number of transcripts and multiplied by 10,000. Scaled expression data were then natural-log transformed before analysis using Seurat[23].

Analyzing species-mixing experiments. In the first experiment, HEK cells were identified as those barcodes with greater than 2,000 human transcripts and less than 1,000 mouse transcripts, while barcodes with greater than 2,000 mouse transcripts and less than 1,000 human transcripts were identified as 3T3 cells. Cells with fewer than 2,000 total transcripts were considered indeterminate, while any cell with greater than 5,000 total transcripts and more than 1,000 nonmouse or nonhuman transcripts was considered a multiplet (**Fig. 1d**). In the second experiment, HEK cells were identified as those barcodes with greater than 10,000 human transcripts, greater than 2,000 human genes, and greater than 90% human transcript alignment; while barcodes with greater than 10,000 mouse transcripts, greater than 2,000 mouse genes, and greater than 90% mouse transcript alignment were identified as 3T3 cells. Cells with fewer than 10,000 total transcripts were considered indeterminate, while any cells with greater than 10,000 total transcripts and more than 1,000 nonmouse or nonhuman transcripts were considered multiples (**Fig. 1c** and **Supplementary Fig. 8**).

PBMC analysis. We reduced the dimensionality of our data to 11 principle components that account for the majority of the variation (51.6% cumulative variance) among variable genes to achieve optimal discrimination of cell types identified through image cytometry. We identified seven distinct clusters of cells using the FindClusters function in Seurat with k.param = 50 (a measure of neighborhood size) and resolution = 0.75 (see below; **Supplementary Fig. 11**). Clusters corresponding to CD4+ T cells, CD8+ T cells, B cells, NK cells, monocytes, and dendritic cells were all identified on the basis of significant enrichment using an ROC test implemented in Seurat (also see **Supplementary Figs. 10** and **11**). We removed 602 cells that comprised a distinct cluster enriched for expression of mitochondrial genes (**Supplementary Fig. 11**) and a lower mapping rate of new transcripts and genes per sequencing read (**Supplementary Fig. 12**), which likely represented single-cell libraries of low complexity. We then applied t-distributed stochastic neighbor embedding (t-SNE) using the cell loadings for the previously chosen 11 principle components to visualize the cells in two dimensions. Following sequence alignment, we analyzed a total of 4,296 cells in which at least 10,000 reads, 1,000 transcripts and 500 genes were detected with mRNA alignment rate greater than 65% (**Fig. 2b–d**), which resulted in filtering of 1,670 cells with greater than 1,000 transcripts. We analyzed a total of 6,713 genes that were detected in at least 2.5% of filtered cells across six sequencing runs from three separate arrays. We identified 687 variable genes with log-mean expression values greater than 0.5 and dispersion (variance/mean) greater than 0.5. We observed optimal discrimination of cell types identified through image cytometry using 11 principal components that account for the majority of the variation (51.6% cumulative variance) among variable genes and visualized using the *t*-distributed stochastic neighbor embedding (t-SNE) algorithm. We performed 1,000 iterations of the Barnes–Hut implementation of the t-SNE algorithm using a 'perplexity' value of 40. We identified seven distinct clusters of cells using the FindClusters function in Seurat with k.param = 50 (a measure of neighborhood size) and resolution = 0.75 (see below; **Supplementary Fig. 11**). Clusters corresponding to CD4+ T cells, CD8+ T cells, B cells, NK cells, monocytes and dendritic Cells were all identified on the basis of significant enrichment using an ROC test implemented in Seurat (also see **Supplementary Figs. 10** and **11**). We removed 602 cells that comprised a distinct cluster enriched for expression of mitochondrial genes (**Supplementary Fig. 11**) and a lower mapping rate of new transcripts and genes per sequencing read (**Supplementary Fig. 12**), which likely represented single-cell libraries of low complexity. We examined proportions of various cell types across arrays and sequencing runs among 3,694 cells that passed the aforementioned filtering criteria. Statistical significance of differences in the proportion of clusters between separate arrays and sequencing runs was performed using a Chi-square test (**Fig. 2c**). We further examined phenotypic variation within myeloid cells among identified principal components (**Fig. 2d**) by ranking cells on the basis of their PC score among genes with highest loadings for each principal component.

Comparison of Seq-Well PBMCs to 10× genomics data. We performed comparisons of gene detection and transcript capture among PBMC cell types conserved between 3,590 PBMCs (excluding dendritic cells) obtained using Seq-Well and 2,700 PBMCs from the 10x Genomics platform (http://support.10xgenomics.com/single-cell/datasets/pbmc3k). To classify PBMC cell types within the 10x Genomics data, we first identified 446 variable genes with log-mean expression values greater than 0.5 and dispersion (variance/mean) greater than 0.5. We then performed graph-based clustering using 13 principal components, k.param of 50 and resolution of 0.75. Cell type identity of each cluster was established on the basis of gene enrichments. Comparisons of genes and transcripts were initially performed between B cells, CD4 T cells, CD8 T cells, monocytes and NK cells using raw data matrices. We refined these comparisons by separately downsampling genes and transcripts within each cell type in Seq-Well data to an average read depth of 69,000 reads per cell to match the reported sequencing depth using in publicly available 10x Genomics data.

***Mycobacterium tuberculosis* analysis.** Following sequence alignment, we identified a total of 14,218 cells with greater than 1,000 mapped transcripts. Initially, we analyzed a subset of 4,638 macrophages with greater than 5,000 detected transcripts (**Supplementary Fig. 14a**) and a total of 9,381 genes expressed

in at least 5% of filtered cells. Principal components analysis was performed among a set of 377 variables genes, defined by genes with log-mean expression greater than 0.5 and dispersion (variance/mean) greater than 0.5. We performed graph-based clustering, as described below, using the first five principal components since we observed that they captured the majority of the biological variation in our data set (63% cumulative variance), and that each additional principal component contributed less than 1% to the total variance. We performed 1,000 iterations of the t-SNE algorithm (Barnes–Hut implementation) using a 'perplexity' value of 30. We identified five distinct clusters of cells in the t-SNE plot using the FindClusters function in Seurat with k.param = 40 and resolution = 0.25 (**Supplementary Fig. 14**). We removed two clusters comprised of cells with reduced gene detection, transcript capture and enrichment for expression of mitochondrial genes. Following removal of low-quality cells, we analyzed three distinct clusters with total of 2,560 high-quality cells (**Fig. 3a** and **Supplementary Fig. 14**). Differential expression analysis was performed between clusters, and cells exposed and unexposed to TB within each t-SNE cluster using a likelihood ratio test in Seurat (**Fig. 3c** and **Supplementary Table 9**). We performed gene set enrichment analysis to examine association of expression differences observed between control macrophages exposed and unexposed to *M. tuberculosis* with previously published gene sets using GSEA. For each cluster, expression patterns between exposed and unexposed cells were made to complete GSEA databases (**Supplementary Tables 7**, **8**, **10** and **11**).

Regressing out latent technical effects. Technical parameters governing sequencing data, such as the number of genes detected or the transcriptomics alignment rate, often vary significantly across single cells. We sought to conservatively remove these technical effects using a 'latent-variable' approach similar to that of Buettner *et al.*[24]. Briefly, we fit a linear model to predict the expression value of each gene based on a set of technical metrics, as well as the total number of unique genes detected in that cell. In our analyses, we constructed models to adjust gene expression values for alignment rate of each cell. We considered the residual expression from this model as a 'corrected' gene expression value, and we used these values as input to the downstream clustering analyses.

Graph-based clustering of single-cell transcriptomes. For all single-cell clustering analyses, we used an approach similar to that of our recently proposed clustering strategy for Drop-seq data. Briefly, as in Macosko *et al.*[12], we first identified the set of genes that was most variable across our data set after controlling for the relationship in single-cell RNA-seq data that inherently exist between mean expression and variability by binning genes into 20 bins based on their average expression level and *z*-scoring dispersion (mean/variance) estimates within a bin. We excluded all genes which were detected in less than 2.5% of PBMCs (5% of monocytes for the Mtb experiments) and used a dispersion cutoff of 0.5 to select variable genes, resulting in the selection of 687 variable genes across 4,296 PBMCs and 377 variable genes across 4,638 macrophages.

We next reduced the dimensionality of our data set, using principal components analysis. As previously described in Macosko *et al.*[12], we ran PCA using the prcomp function in R. We then selected PCs for further downstream analysis (11 PCs in PBMC analysis and 5 PCs in TB analysis). As expected, markers for distinct cell types were highly represented among the genes with the largest scores along these PCs. We then applied *t*-distributed stochastic neighbor embedding (t-SNE) using cell loadings for the significant principal components as input to visualize the structure of our data in two dimensions.

Here we used graph-based clustering methods, similar to those that have been recently proposed for both single-cell RNA-seq and mass cytometry data[25,26]. We first construct a Euclidean distance matrix on the loadings for the significant principal components as described above and use this to construct a *K*-nearest neighbor graph (KNN, $k = 50$ in PBMC analysis; $k = 40$ in TB analysis). Our goal was to identify 'quasi-cliques'[26], or 'communities'[25], of cells that were highly interconnected across this graph. Therefore, we first converted the KNN graph into a weighted shared nearest neighbor (SNN) graph, where the weight between any two cells was represented by the percent overlap in their respective *K*-nearest neighborhoods (Jaccard distance), and we pruned low-quality edges with a Jaccard distance of <0.1 (less than 10% overlap in local neighborhoods). Finally, to group the cells into clusters, we used a recently developed method for modularity optimization, which aims to optimize a function describing the density of connections within a cluster versus connections between clusters, essentially to identify highly interconnected nodes within the SNN graph. Here, we applied the smart local moving algorithm, which is similar to the widely used 'Louvain' algorithm for community detection but implements a local moving heuristic that enables communities to be split up and iteratively reorganized in an attempt to improve the overall partition modularity. This grants the SLM algorithm additional freedom in identifying an optimal clustering solution, and we empirically observed increased sensitivity and consistency applying this approach to single-cell data.

Data availability. All RNA-seq data are available in GEO under accession number GSE92495. Source data files for **Figures 1–3** are availabel online.

21. Gierahn, T. *et al.* Seq-Well protocol. *Protocol Exchange* http://doi.dx.org/10.1038/protex.2017.006 (2016).
22. Roederer, M. *Current Protocols in Cytometry* (eds. Robinson, J.P. *et al.*) Unit 1.14 (Wiley, 2002).
23. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. *Nat. Biotechnol.* **33**, 495–502 (2015).
24. Buettner, F. *et al. Nat. Biotechnol.* **33**, 155–160 (2015).
25. Levine, J.H. *et al. Cell* **162**, 184–197 (2015).
26. Xu, C. & Su, Z. *Bioinformatics* **31**, 1974–1980 (2015).

doi:10.1038/nmeth.4179