Supplementary Material for

Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay,* Ido Amit*

‡Corresponding author. E-mail: amos.tanay@weizmann.ac.il (AT); ido.amit@weizmann.ac.il (IA)

Published 14 February 2014, *Science* **343**, 776 (2014)
DOI: 10.1126/science.1247651

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/343/6172/776/suppl/DC1)

Tables S1 to S6 as separate Excel files

MATERIALS AND METHODS

Isolation of splenic CD11c+ cell suspension

Spleens were extracted from C57BL/6J female mice (8 to 12 weeks old), dissociated into single splenocytes with a gentleMACS Dissociator (Miltenyi Biotec, Germany) and incubated 5 min in red blood cell lysis solution (Sigma). Cells were then washed and resuspended in MACS buffer (0.5% BSA and 2 mM EDTA in phosphate-buffered saline), and filtered through a 70-μm strainer. A $CD11c^+$ fraction was obtained through two rounds (double-enrichment) of separation with monoclonal anti-mouse CD11c antibodies coupled to magnetic beads using a MACS cell separator system (Miltenyi Biotec).

Single cell capture

Single cells were sorted into cell capture plates, containing 2 μl of cell lysis solution in 384-well PCR plates. Capture plates were prepared with a Bravo automated liquid handling platform (Agilent). Sorting was performed using a FACSAria III cell sorter (BD Biosciences) and gating in SSC-A vs. FSC-A to collect live cells, and then in FSC-W vs. FSC-A to sort only singlets. Two empty wells were kept in each 384-well plate, as a no-cell control during data analysis. Immediately after sorting, plates were spun down to ensure cell immersion into the lysis solution, snap frozen on dry ice and stored at -80°C until further processing.

MARS-Seq: 384-well plates automation setup and library construction

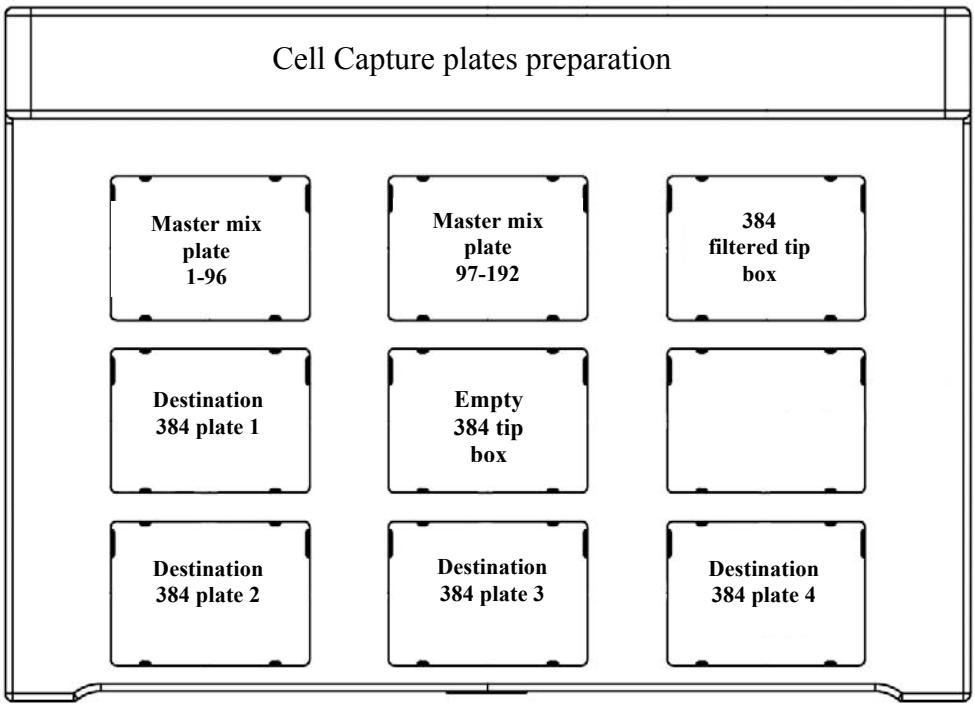
Automated single-cell RNA-Seq library production was performed on the Bravo robot station using 384-filtered tip (Axygen, catalog # 302-82-101). The Bravo Single-cell RNA-Seq scripts are available upon request. The samples were processed as described in the following steps:

1. <u>384-well cell-capture plate preparation protocol</u>

Preparation of 384-well single-cell capture plates included the addition of 2 μl of a hypotonic cell lysis solution (a robust splenic lysis solution compatible with a cell direct RT reaction) supplemented with RNase inhibitor and a barcoded RT primer. The RT primer included a T7 RNA polymerase promoter, a partial Illumina paired-end primer sequence, a cell barcode followed by a unique molecular identifier (*23*), and an anchored polydT (see table S7):

1. 96-well master mix plates contain lysis solution (triton 0.2% in molecular biology-grade water) supplemented with 0.4 U/μl RNase inhibitor and 400 nM of indexed RT primer from group 1 (1-96 barcodes) or group 2 (97-192 barcodes). To prepare 12 384-

well plates, 57.5 µl lysis buffer were mixed with 5 µl of 5 µM indexed RT primer stock per well.
2. The cell capture plate preparation script mixes group 1 master mix plate (barcodes 1-96), aspirates 2 µl from it and dispenses it in destination 384-well plate-1 in two adjacent positions (see below). Then, 2 µl are again aspirated from master mix plate 1-96 to be dispensed in the other destination 384-well plates. If more than four cell capture plates are needed, filled destination plates should be replaced with new plates. Once all desired plates are added with group 1 master mix, tips are replaced and the cell capture plate preparation script mixes group 2 master plate (barcode 97-192), aspirates 2 µl from it and dispenses it in the destination 384-well plates. The entire process takes about 30 min per 12 plates. A single cell is then sorted into each well as described above.

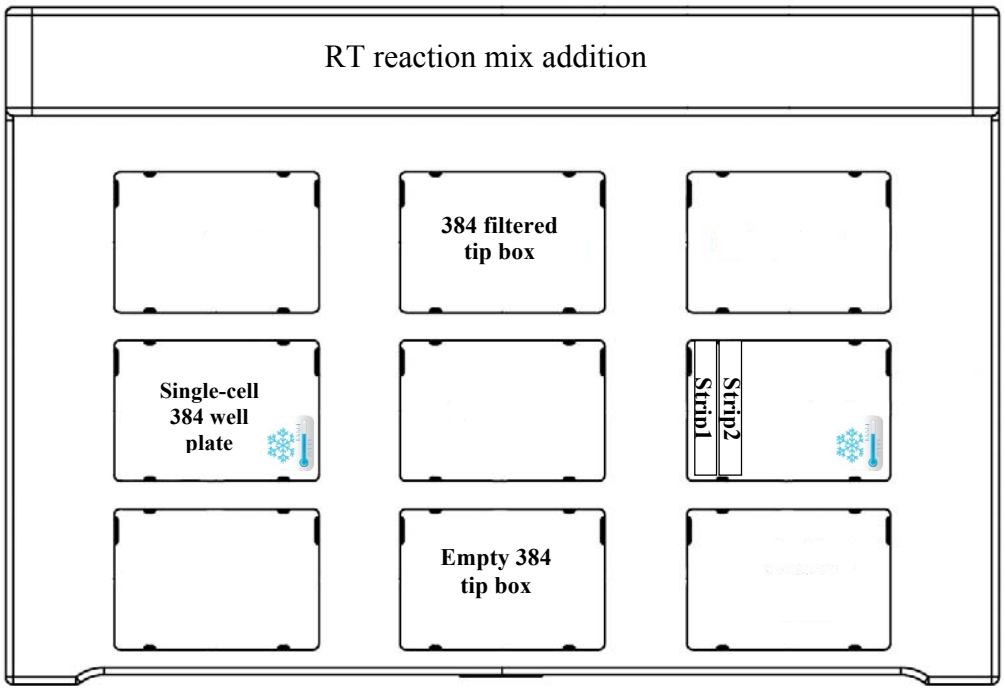


2. Barcoding and Reverse Transcription (RT) reaction

1. Pre-incubation: To open secondary RNA structures and allow annealing of the RT primer, the 384-well cell capture plate was incubated at 72˚C for 3 min and immediately transferred to 384-well Inheco thermal block integrated to Bravo and set at 4˚C (see position 4 at Bravo RT protocol scheme below).

2. An RT reaction mix (10 mM DTT, 4 mM dNTP, 2.5 U/µl Superscript III RT enzyme in 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$, ERCC RNA Spike-In mix) was prepared as a master mix for 440 reactions (sufficient for 384 wells and the required void volume for automatic pipetting), dispensed into two 8-well strips, 54 µl per well, and then placed in a 96-well Inheco thermal block set at 4˚C (position 6 in Bravo RT scheme below). The RT reaction mix was supplemented with ERCC (*24*) RNA Spike-In mix

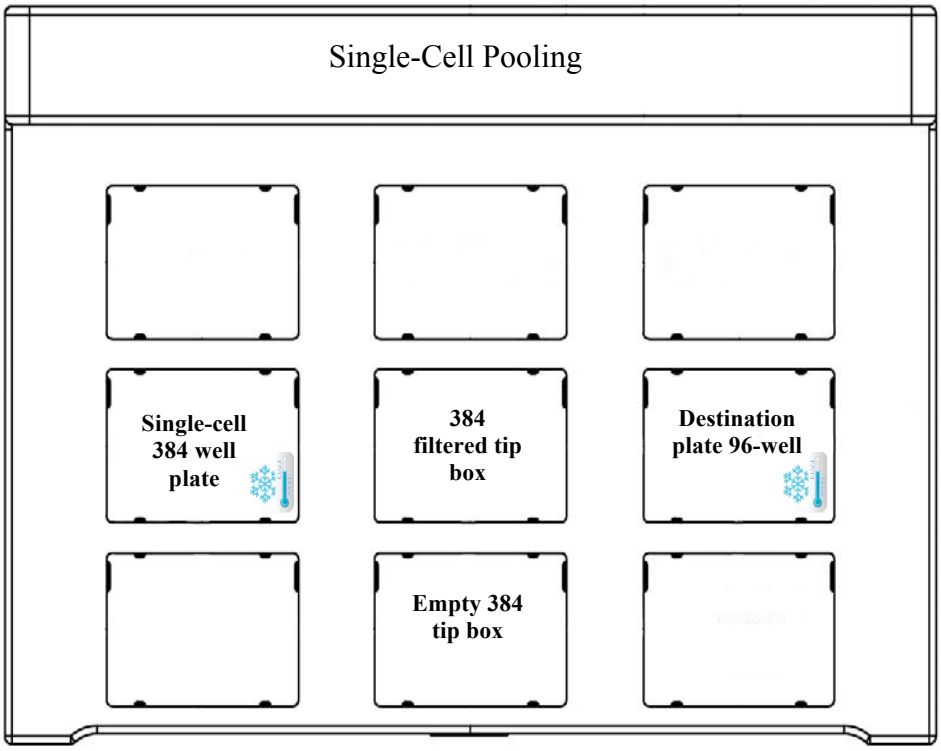
(Ambion), containing polyadenylated RNA molecules of known length and concentration, at a final 1:40x10^7 dilution per cell, to yield ~ 5% of the single-cell mRNA content.
3. According to the RT reaction mix addition script, 2 µl of RT reaction mix were added into each well of the 384-well plate (described in step 1) and the reaction was mixed one time. Tips were replaced and the process repeated to all wells.
4. The 384-well plate was then spun down and moved into a 384 cycler (Eppendorf) for the RT program: 2 min at 42˚C, 50 min at 50˚C, 5 min at 85˚C. The entire process takes 23 min per 384-well plate.



3. <u>Pooling of barcoded 384 single cell samples</u>

1. Tips pre-wash and blocking: 50 µl of triton 0.2% containing 40 pg/µl of yeast tRNA were dispensed in each well of row D of a clean 96-well plate (destination plate for pooling).
2. The 4 µl of barcoded cDNA sample from the 384-well plate (placed on the 384-well Inheco block set at 4˚C) were pooled into two rows (24 wells) of the 96-well destination plate (placed on the 96-well Inheco block set at 4˚C). The entire process takes 9 min per plate.
3. To remove RT primer leftovers, 1 µl of exonuclease I (NEB) was added into each of the 24 wells and the plate incubated at 37˚C for 30 min and then 10 min at 80˚C for exonuclease I inactivation.
4. A second and final pooling was achieved through sample cleanup, consisting of addition of 1.2x volumes of SPRI magnetic beads into each well (low SPRI to sample ratio further removes primer traces) to bind the cDNA, and the contents of each row

(containing 192 barcoded single cells) are pooled into one 1.7 mL DNA low-bind Eppendorf tube, washed and eluted in 17 µl 10 mM Tris HCl pH 7.5. These two groups of 192 samples will be pooled together (along with cells from three other 384-well plates) after addition of a pool barcode (plate level tag; see S8) in the RNA-DNA ligation step (see below).



4. Second strand synthesis and IVT amplification

1. The pooled cDNA was converted to double-stranded DNA with a second strand synthesis kit (NEB) in a 20 µl reaction, incubating for 2.5 h at 16˚C.
2. The product was purified with 1.4x volumes of SPRI beads, eluted in 8 µl and in-vitro transcribed (with the beads) at 37˚C overnight for linear amplification using the T7 High Yield RNA polymerase IVT kit (NEB).
3. Following IVT, the DNA template was removed with Turbo DNase I (Ambion) 15 min at 37˚C and the amplified RNA (aRNA) purified with 1.2x volumes of SPRI beads.

5. Single-cell library preparation for high-throughput sequencing

1. The aRNA was chemically fragmented into short molecules (median size ~200 nucleotides) by incubating 3 min at 70˚C in Zn^{2+} RNA fragmentation solution (Ambion) and purified with two volumes of SPRI beads.
2. Next, a partial Illumina Read1 sequencing adapter that includes a pool barcode (see table S8) was single strand ligated (ligation adapter sequence is in table S7) to the

fragmented RNA using a T4 RNA ligase I (New England Biolabs). The aRNA (5 µl) was preincubated 3 min at 70˚C with 1 µl of 100 µM ligation adapter; then, 14 µl of a mix containing 9.5% DMSO, 1 mM ATP, 20% PEG8000 and 1 U/µl T4 ligase in 50 mM Tris HCl pH7.5, 10 mM $MgCl_2$ and 1mM DTT was added. The reaction was incubated at 22˚C for 2 h.
3. The ligated product was reverse transcribed using Affinity Script RT enzyme (Agilent; reaction mix contains Affinity Script RT buffer, 10 mM DTT, 4 mM dNTP, 2.5 U/µl RT enzyme) and a primer complementary to the ligated adapter (see table S7). The reaction was incubated for 2 min at 42˚C, 45 min at 50˚C, 5 min at 85˚C. The cDNA was purified with 1.5x volumes of SPRI beads.
4. The library was completed and amplified through a nested PCR reaction with 0.5 µM of each primer and PCR ready mix (Kapa Biosystems). The forward primer contains the Illumina P5-Read1 sequences and the reverse primer contains the P7-Read2 sequences (see table S7). The amplified pooled single-cell library was purified with 0.7x volumes of SPRI beads to remove primer leftovers. Library Concentration is measured with a Qubit fluorometer (Life Technologies) and mean molecule size is determined with a 2200 TapeStation instrument (Agilent Technologies). MARS-Seq libraries were paired-end sequenced using an Illumina HiSeq 2500. We sequenced from 192 to 1,536 cells per lane.

Single cell sorting efficiency assessment

To test single cell sorting efficiency we used two parallel approaches. First, we directly sorted single cells and measured cell frequency using a semi-automated image analysis software (fig. S3). Second, during single-cell sorting we designated one well to remain empty to be used for no-cell background signal during analysis.
For image analysis, cells were stained with carboxyfluorescein succinimidyl ester (CFSE; eBioscience, San Diego, CA) and sorted into two 96-well plates and scanned using a fluorescence microscope. Briefly, HEK 293T cells from an exponential culture were harvested and stained with 2.5 µM CFSE in PBS for 10 min, then washed three times with seven volumes of complete medium (10% serum; no phenol red). After sorting single cells into wells containing 100 µl of culture medium, the plates were spun down and scanned using a Ti-eclipse microscope (Nikon Instruments, Melville, NY) equipped with an automated stage, an incubator, and a closed chamber that allows for CO_2 flow over the 96-well plate. For cell detection calibration, we sorted 10 or 100 cells into the wells of column 1 of each plate. Cells were imaged using a 10X objective and monitored using bright field illumination and fluorescence channel FITC. Pictures were collected using the NIS-elements software (Nikon Instruments). For quantitative analysis, images from the fluorescence channel were analyzed using a semi-automated ImageJ (National Institutes of Health, Bethesda, MD; available at http://rsb.info.nih.gov/ij/index.html) based custom code. Wells that were not properly scanned by the microscope were excluded from the analysis.

We observed that 2.3% of wells contained no cells, 2.3% of wells contained 2 cells, and more than 95% with one single cell. We did not detect any well with more than two cells. Of note, in all wells with two cells, these cells were always next to each other, raising the possibility that they originated from one dividing cell that continued to divide

after having been sorted into culture medium. Since in the single-cell RNA-Seq process cells are sorted into a hypotonic solution, these numbers probably represent an overestimation of the true doublets we have in our data.

Isolation of DC subpopulations by Fluorescence-activated cell sorting

For sorting DC subpopulations, MACS-based CD11c enriched mouse splenocytes were stained and sorted on a FACSAria III cell sorter (BD Biosciences) in two rounds, using fluorophore-conjugated antibodies (BioLegend). First, cells were stained with FITC-conjugated anti-CD8a antibodies (clone 53-6.7) and sorted into CD8a positive and negative fractions. The $CD8^+$ fraction was then stained with APC anti-CD11c (clone N418), Pacific Blue anti-MHCII (clone AF6-120.1), Alexa 700 anti-CD4 (clone GK1.5), PE-Cy7 anti-CD86 (clone GL-1), and PE-conjugated anti-PDCA1. The $CD8^-$ fraction was stained for CD11c, MHCII, and with PerCP-Cy5.5 anti-CD11b, PE-Cy7 anti-CD4, FITC anti-PDCA1, and PE-conjugated anti-ESAM (clone 1G8). The DC cells were identified as: cDC $CD8^+$ ($CD11c^{high}$ $MHCII^+$ $CD8a^{high}$ $CD86^+$); cDC $CD86^-$ ($CD11c^{high}$ $MHCII^+$ $CD8a^{inter}$ $CD86^-$); $CD8^+$ pDC ($CD11c^{inter}$ $CD8a^+$ $PDCA1^+$); cDC $CD4^+$ $ESAM^+$ ($CD8^-$ $MHCII^+$ $CB11b^+$ $CD4^+$ $ESAM^+$); $CD8^-$ pDC ($CD11c^{inter}$ $CD8a^-$ $PDCA1^+$). For single-cell sequencing, single cells were sorted into 96/384 well single-cell capture plates as described above.

Isolation of hematopoietic cell types

To obtain B cells, NK cells and monocytes, a splenocyte suspension was stained with, PE-Cy7-conjugated CD19, eFluor 450-conjugated NK-1.1, PerCP Cy5.5 Gr1, FITC TCR-β, APC CD11b and PE B220 (CD45R). $B220^+$ and $B220^{neg}$ (germinal center) B cells were collected by gating for $CD19^+$ (TCR-β^{neg}) cells and then by B220 against the CD19 marker. NK single cells were collected from the CD19neg/TCR-β^{neg} events by gating for NK-1.1 positive events in NK-1.1 vs. Gr1. Finally single monocytes were collected by gating for $Gr1^+$ $CD11b^+$ events. The B cell and pDC content in the CD11c-enriched sample was estimated by staining with PE-Cy7 CD19, PE PDCA-1 (CD317, Bst2) and APC CD11c and gating in CD19 vs. CD11c and PDCA-1 vs. CD11c, respectively. For single-cell sequencing, single cells were sorted into 96/384-well single-cell capture plates as described above.

Single-cell Real Time PCR

Single B cells, NK cells and monocytes were sorted by FACS into individual wells of a 96-well plate containing 5 µl of 0.2% Triton X-100 and RNase inhibitor as described above. RT pre-amplification was performed on 24 single cells of each type similarly to Dalerba, et al. (*25*). After thawing, each well was supplemented with 0.1 µl of SuperScript III RT/Platinum Taq (Invitrogen), 6 µl of 2x reaction mix and a mixture of primer pairs for *Cd37* (B cell marker), *Ly6A* (B cell marker), *NKg7* (NK marker) and

Ccl4 (NK cell marker) genes (100 nM final concentration; see table S9 for sequences). Single-cell mRNA was directly reverse transcribed into cDNA (50°C for 15 min, 95°C for 2 min), preamplified for 14 cycles (each cycle 95°C for 15 sec, 60°C for 1 min) and cooled at 4°C for 15 min. Samples were then diluted 1:40 with 10 mM Tris-HCl, pH 8. Real- Time PCR analysis was performed for each gene separately with the same set of primers used in the RT pre-amplification stage (400 nM final concentration) using SYBR green Master (Roche) on a LightCycler 480 System instrument (Roche). Quantification was performed as relative to the average of all cells for a given gene (n = 72), using the formula 2(Ct – mean (Ct)), where Ct is the mean qPCR cycle threshold signal of two replicate qPCR reactions per cell.

Structure of valid library products and their expected distributions

Following the final amplification, single-cell RNA-seq sequenced product information was structured in two parts. At one end (Read 1) we read a 56-57 bp sequence that included a 6 bp pool barcode prefix (table S8) followed by a sequence expected to map within a polyadenylated transcript. For valid library products, this fragment is expected to map at some typical (short) offset from the gene's 3' UTR, depending on the randomized fragmentation of the IVT-amplified RNA (library construction protocol; see above). The other sequence end (Read 2) contains a 10-14 bp tag that was engineered to include a 6 bp cell-specific (or well-specific) label, followed by a 4-8 bp random molecular tag (RMT).

Importantly:
1. Groups of reads that share a pool-barcode, cellular tag and RMT were assumed to represent the same initial RNA molecule and were counted only once. Such reads typically map to several positions around the 3'UTR of the gene, since multiple IVT products sharing the same tag are fragmented at different offsets.
2. Our pool barcodes and cell-specific labels are designed to be distinct enough (in terms of edit distance) so to reduce the probability of inter-cell contamination through sequencing error. RMTs, on the other hand, are distributed randomly (and unevenly) over all possible DNA k-mers, making sequencing errors difficult to detect or correct (see below).
3. When deep-sequencing a single-cell library, we expect a variable number of reads to cover each RMT. The sequencing depth per molecule mostly depends on its ligation yield and PCR efficiency, which are expected to be similar between molecules that map to the same genomic position. We therefore expect molecules representing the same gene and same 3' UTR offset to be covered relatively uniformly and can use such uniformity assumption for normalization.
4. RMTs mark unique molecules with high probability. However the probability of observing two distinct molecules labeled by the same RMT is not zero, especially for genes that are highly expressed. RMTs of 8bp reduce this effect considerably, but our current dataset include almost only experiments using 4bp RMTs.

Initial filtering, tag extraction and mRNA sequence mapping

Given raw sequenced reads, we first extract pool barcode, cell-specific tags and RMTs, and eliminate reads with ambiguous plate/cell-specific tags or RMT sequence with low quality (Phred<27). We filtered potentially bacterially originated molecules by mapping R1 reads to e-coli using bowtie with parameters "-M 1 -t --best --chunkmbs 64 –strata". Following this initial filtering, we mapped R1 reads (trimmed of pool barcode) to the mouse mm9 genome assembly combined with ERCC spike-in pseudo-assembly using the Bowtie program and the standard parameters "-m 1 -t --best --chunkmbs 64 –strata".

We define a set of transcription termination sites (TTS) based on the UCSC genome browser tables (mm9). Sequence reads mapping to a range of -1000 to +200 bp from a known TTS were considered for further analysis. This leaves out of the analysis less than 20% of the sequenced products, likely representing non-classical genes, alternative 3'UTRs, or spurious transcripts. Following this procedure, we generated a matrix containing the number of reads covering each of the RMTs in each of the observed mapping offsets for each cell and each gene. In ambiguous cases, when reads could be mapped by this procedure to multiple genes, we added a new gene-like record, defined as an underscore delimited list of all these genes. This matrix is then further processed to eliminate biases and errors.

Filtering RMT sequencing errors

As outlined above, sequencing errors introduced within random Molecular Tags (RMT) in our library products may undermine the tag-counting approach by creating spuriously identified molecules from real molecules. The number of such spurious RMTs is expected to scale linearly with the number of times each real RMT is sequenced. However, RMT sequencing errors are incapable of changing the offset of the mapped read relative to the TTS, and for each spurious RMT we expect to identify the *source* RMT as a highly covered tag sharing all the offsets of the spurious RMT (see examples in fig. S4).

Based on these assumptions we implemented the following greedy filtering procedure, applied separately for the set of reads assigned to a certain gene/cell pair:
* Sort the RMTs given their number of unique mapping offsets
* Repeatedly selecting the RMT **T** observed at the fewest offsets, and testing if there exist a *source RMT* **S**, which is a) observed at all the offsets of **T** and b) has an edit distance of 1 from **T**. If such a source RMT exists, we eliminate T and its associated reads.

While cell-barcodes are generally more robust to sequencing errors (given their design), we are identifying potential sequencing errors leading to cell-barcode mismatch using an approach similar to the one described above for RMTs by analyzing sets of molecules with the same RMT but different cell-barcodes and testing if the molecules within one cell are dominated by molecules in another cell with a poorly separated barcode.

Identifying and filtering skewed offsets and cross-cell contaminations

Minimizing cross-cell contamination is important for any single-cell RNA-seq pipeline, but it becomes particularly critical when scaling up the protocol to a large number of cells or when applying it to a heterogeneous sample. Even relatively small levels of read to cell association errors can create a strong background and batch effect, increase spurious correlations between cells and reduce the capability of the approach to detect small coherent subpopulations. In theory, contamination is prevented by well-specific labeling, since the label is retained after pooling the tagged cDNA from all cells and the subsequent stages in the pipeline. Nevertheless, the extensive PCR amplification performed during library construction, and the existence of common (poly-dT) sequences at one end of the library products may give rise to unexpected scenarios of "tag-switching" and read mislabeling. We therefore studied the complex distributions of reads over cells, genes, 3'UTR offsets, and RMTs in our data, aiming to identify and eliminate such potential noise factors. We discovered that there exist certain genomic positions that show statistically unexpected high frequency of low coverage molecules, which should be preferably filtered to minimize different biases.

We implement a filtering of such problematic genomic positions in the following way. Given a set of reads that map onto a given gene, we first define the set of *U* triplets *(c,o,T)* with c being the cell, *o* being the genomic position and *T* being the RMT. We define the set of *lonely* triplets U_l as all (c,o,T) such that there does not exist o' with (c,o',T) in U other than o. The *friendly* triplets U_f are defined as $U-U_l$. We also define the sets U(o) as the triplets in a given position o. We can now compute how statistically unlikely is it to find many lonely triplets at a certain position by commuting the hyper-geometric p-value:

$H(|U|, |U_l|, |U(o)|, intersect(|U_l|, U(o)))$

We perform Benjamini-Hochberg FDR correction on these p-values and exclude all triplets at offsets that obtain FDR<0.25. Further filtering is done using the same procedure, but defining the set of lonely triplets as those covered by only one read.

We reasoned that genomic positions with significant hyper-geometric p-value are likely to be prone to barcode-switching effects, and case-by-case analysis confirmed that many of these are giving rise to batch-specific effects and are important to eliminate. Thus we recommend diagnosing and filtering such effects in single-cell RNA-seq studies.

Down-sampling normalization

Unlike any other gene expression datasets, single-cell mRNA profiles are inherently discrete, representing limited sampling from the initially limited pool of mRNA molecules within each cell. The number of trustworthy sampled molecules per cell is

variable and in order to compare profiles between cells, normalization is sometime desirable. Since we demonstrated our samples can be considered as sparse multinomial samples of mRNAs from each cell (e.g., Figure 1), the only appropriate normalization scheme is probabilistic: we define a target number of molecules N, and then sample from each cell having m>=N molecules precisely *N* molecules without replacement. Cells with m<N are not used for analysis at this level. This down-sampling approach ensures that all normalized cells should reflect the same family of multinomial distributions and can be robustly compared. We note that common practices in normalization of gene expression data (e.g. dividing by mean or median) must be avoided in single-cell RNA-seq datasets as they introduce severe coverage biases to the analysis.

A multinomial mixture model for single-cell RNA-seq data

A typical single-cell mRNA sample is defined by a vector n_j (number of molecules observed for gene j), measured of a set of cells K. We assume this vector represents the results of sampling from a pool of few hundred thousands mRNAs within each cell, and that this original pool is different from cell to cell depending on the cell type or regulatory state. To model the sampling and cell-type component of the cell-to-cell variability (but not the cell state component) we introduce the following simple multinomial mixture model. The model probabilistically generates vectors of mRNA molecule counts over some space of genes *G*. We assume some number of classes *K* where each class defines a different multinomial distribution over the genes *G*, denoted $p_{ij}=Pr(g_j \mid class=i)$ (the probability of sampling gene j given that we are in mixture i). We also define for each class a mixture coefficient a_i. Given a single-cell data vector n_j we can compute the log probability of the data given each of the class by simple summation: $log(Pr(n_j \mid class\ i)) = \Sigma_j\ n_j\ log(p_{ij})$. Classification of a cell can now be done by comparing these log likelihoods (either choosing the maximum *a posteriori* class, or computing posterior probabilities).

Inference of the mixture model parameters – non-iterative and iterative approaches

We used the following simple, non-iterative approach to infer model parameters in highly heterogeneous, multi-type populations such as those studied in Figure 2 of the main text:

Algorithm 1:
1. We down-sample all data to N=600
2. We compute the mean and variance for all genes in the dataset, and select the 100 genes with the highest variance/mean ratio, considering only genes with mean expression within the range (1e-4, 0.014).
3. We perform hierarchical clustering on the single-cell profiles, using the high variance genes.
4. High correlation sub-trees are identified as *seed*s and sets of cells within each seed are extracted (denoted E_i**).** In Figure 2, Seed 7 was defined as union of several poorly separated seeds.

5. We pool together all molecules from cells in the seed E_i, generating a seed count vector m'^i_j. Next we normalize this vector by down-sampling a fixed number of molecules N_{seed} without replacement to create m^i_j. The multinomial parameters are now estimated as:

$$p_{ij} = (m^i_j + k_{reg})/(N_{seed} + k_{reg}\,|G|)$$

where k_{reg} is a regularization constant, |G|, is the number of genes. In the analysis described by Figure 2, N_{seed}=13800 and k_{reg} was set to 1.

We used this simple, non-iterative approach in our initial proof of concept experiments, in which clear cell type hierarchy is evident and require little computational fine-tuning to substantiate. We suggest this approach is relatively easy to understand and interpret, even by non-experts. Analysis of more challenging cell populations can be problematic using this scheme, and may require more sensitive methods that can identify structure in the data even when individual gene count vectors are very sparse. We approach this (**Fig S17**) using an EM-like approach consisting of a greedy initialization step followed by model update iterations:

Algorithm 2:
Preprocess: we used here all cells with a minimal number of molecules (N>400) and all genes with a minimal number of total molecules across all cells (N>40). We did not down-sample the data.
1. We select a first seed cell randomly (denoted k_1)
2. We initialize a multinomial model from the new seed (as discussed above: $p_{ij} = (n^{k1}_j + k_{reg})/(\Sigma_j\, n^{k1}_j + k_{reg}\,|G|)$, and compute the likelihood of all cells given the new model ($\Sigma_j\, n_j\, log(p_{ij})$), divided by the number of molecules ($\Sigma_j\, n_j$) (since data is not down-sampled normalized in this algorithm).
3. We join the D cells with the highest normalized likelihood for the seed model (D=20) and reinitialize the seed model using data from these D cells.
4. We now re-compute the likelihood of each cell to the current model (which include all seeds selected so far).
5. We sample a new seed cell randomly from the set of cells whose likelihoods for the current model are within the bottom 20 percentiles.
6. We continue with steps 2-5 until we have a model with K seeds. We are assuming uniform mixture coefficient throughout this process.
7. Given the now complete model, we recomputed likelihoods of each cell to each seed
8. We determine the maximum *a posteriori* (MAP) class for each cell (again assuming uniform mixture coefficients), and reinitialize the multinomial parameters of each class given the MAP cell-to-class assignment.
9. We terminate iterations on 7-8 once the MAP assignment converges, or once a maximal number of iterations were performed.

Circular *a posteriori* projection (CAP-) visualization

Given a mixture model (as constructed by algorithm 1 or algorithm 2), and a collection of (non down-sampled) single-cell profiles n_j^k (specifying the number of molecules for gene j in cell k), we can compute the probability of the data for each class

$U_{ik} = Pr(n^k \mid class=i)$.

We standardize these values given the total molecules in each sample in order to avoid introducing a coverage bias and normalize over all classes i to generate a corrected "posterior" probability:

$u'_{ik} = exp((100/sum_j(n_j^k) * log(u_{ik}))/Z_k$

Here Z_k is a normalization factor computed so that the values for cell k sum up to 1. To visualize this high dimensional data we define a *circular projection* by assigning each class with a radial position a_i on the unit circle, and assigning each cell with the coordinates:

$x_k = sum_i (u'_{ik} cos(a_i))$, $y_k = sum_j (u'_{ik} sin(a_i))$.

Radial positions are selected to minimize the inconsistencies for cells with ambiguous class posteriors u'_{ik}. Specifically, pairs of classes with many cells mapping ambiguously to them should be positioned on proximal radial positions. To find an assignment of radial positions given these goals we construct a complete graph over the classes, and solve a traveling salesman problem over this graph with distances that represent the inverse number of cells with strong joint posterior probability for each pair of classes. Specifically we compute the joint posterior matrix by multiplying the matrices U' (as defined above):

$V=U'U'^T$

We normalize the product:

$v'_{ii'} = v_{ii'} * (1/sum_l(u_{il}) * sum_l(u_{i'l}))$

and generate a distance matrix:

$d_{ii'} = exp(-10v'_{ii'})$.

Solving the TSP problem for small graphs such as ours is easy to achieve by exhaustive enumeration over all permutations or using one of the standard R packages for TSP. Given an optimal tour on the graph, we assign the radial positions proportionally give the distances on the tour.

Pooling subpopulations for gene clustering

Given a mixture model we classify cells based on their likelihood scores as described above, and associate each cell with its maximum *a posteriory* class. We then pool

together all cells associated with a class and generate a combined vector of RNA counts. We normalize this vector by simple scaling and regularization: $E_j = log(0.0001 + n_j/\Sigma_j n_j)$ (note that here we are not down-sampling the data anymore). We use the resulting gene expression estimation to cluster genes as shown in e.g., Figure 3 and Figure S12.

Testing subpopulation differential expression and comparison to ImmGen

To test differential gene expression between groups of single cells, we performed a standard chi-square-based proportion test on genes with overall mean expression generating at least 7 expected molecules per class. We corrected p-values for multiple testing using Benjamini Hochberg procedure (FDR<0.05). For genes with lower mean expression values, we used a fisher exact test for comparing each of the groups to the complement set.

To compare our data to previously established microarray-based gene expression signatures from the ImmGen project, we quantile-normalized each microarray profile to the distribution of overall molecule counts on the pool of all $Cd11c^+$ single-cell profiles. We then computed the likelihood of the normalized ImmGen profile to each of the mixture model classes, and reported the highest likelihood matches in Figure 2. In figure 2C we include only data on ImmGen profiles that were ranked among the three highest likelihood matches for at least one of the mixture classes.

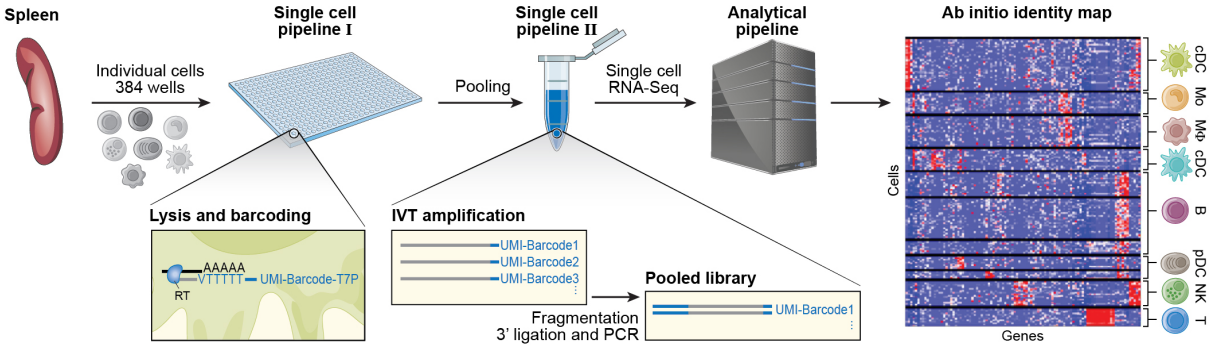


Figure S1. **Massively parallel RNA single-cell sequencing framework (MARS-Seq)**. Schematic diagram of the massively parallel approach to single-cell RNA-seq, involving the use of randomized molecular tags to initially label poly-A tailed RNA molecules, followed by pooling labeled samples and performing two rounds of amplification, generating sequencing ready material (see fig. S2 and methods for an expanded version).

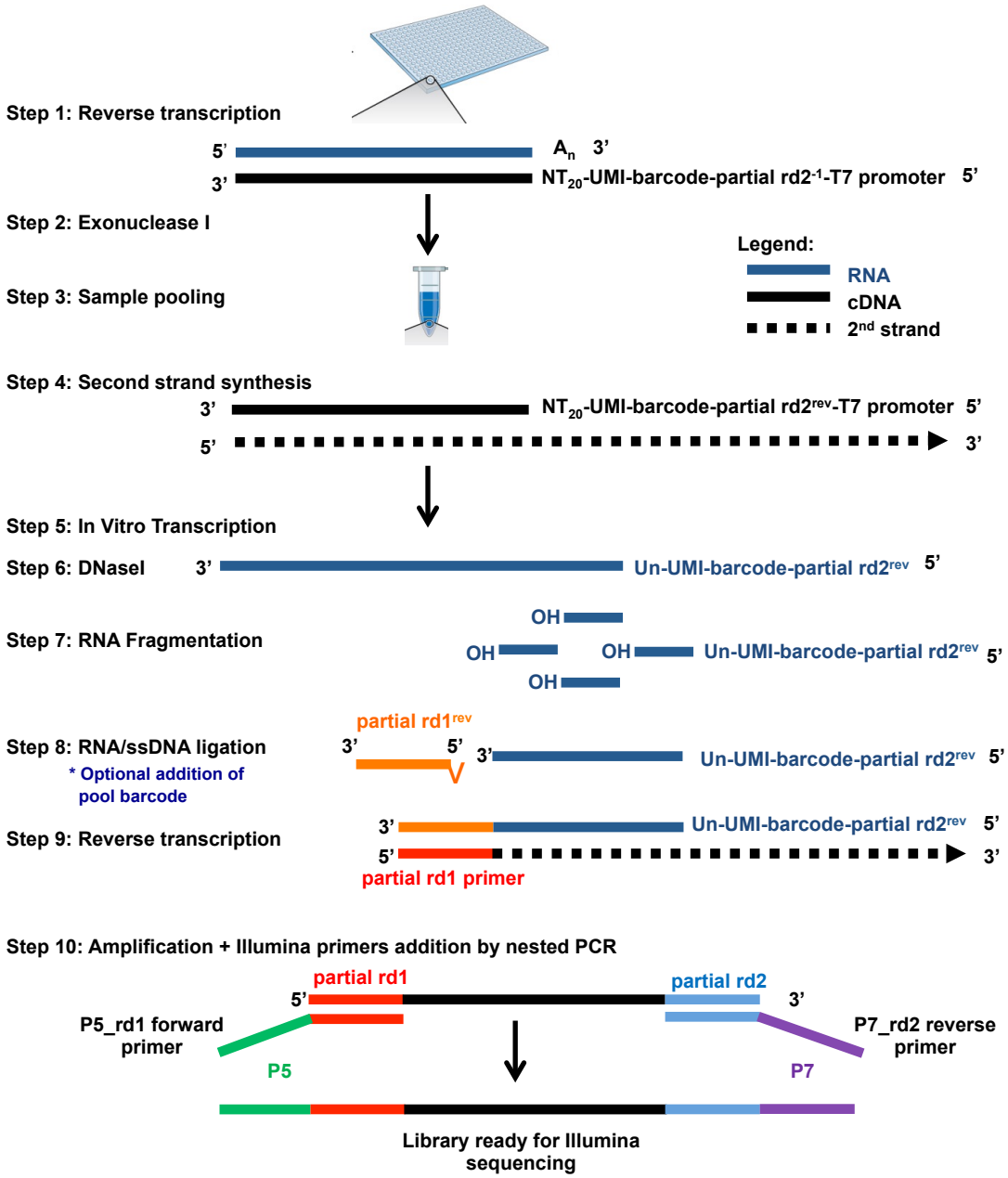


Figure S2. Experimental procedure. Schematic diagram presenting the process of converting single-cell RNA samples to sequencing-ready DNA libraries. Shown are ten experimental steps describing how RNA is tagged, pooled, amplified, fragmented, and how library construction is being performed. Colored lines represent RNA (blue) or DNA (black) molecules, or oligos and primers (see methods for a detailed description).

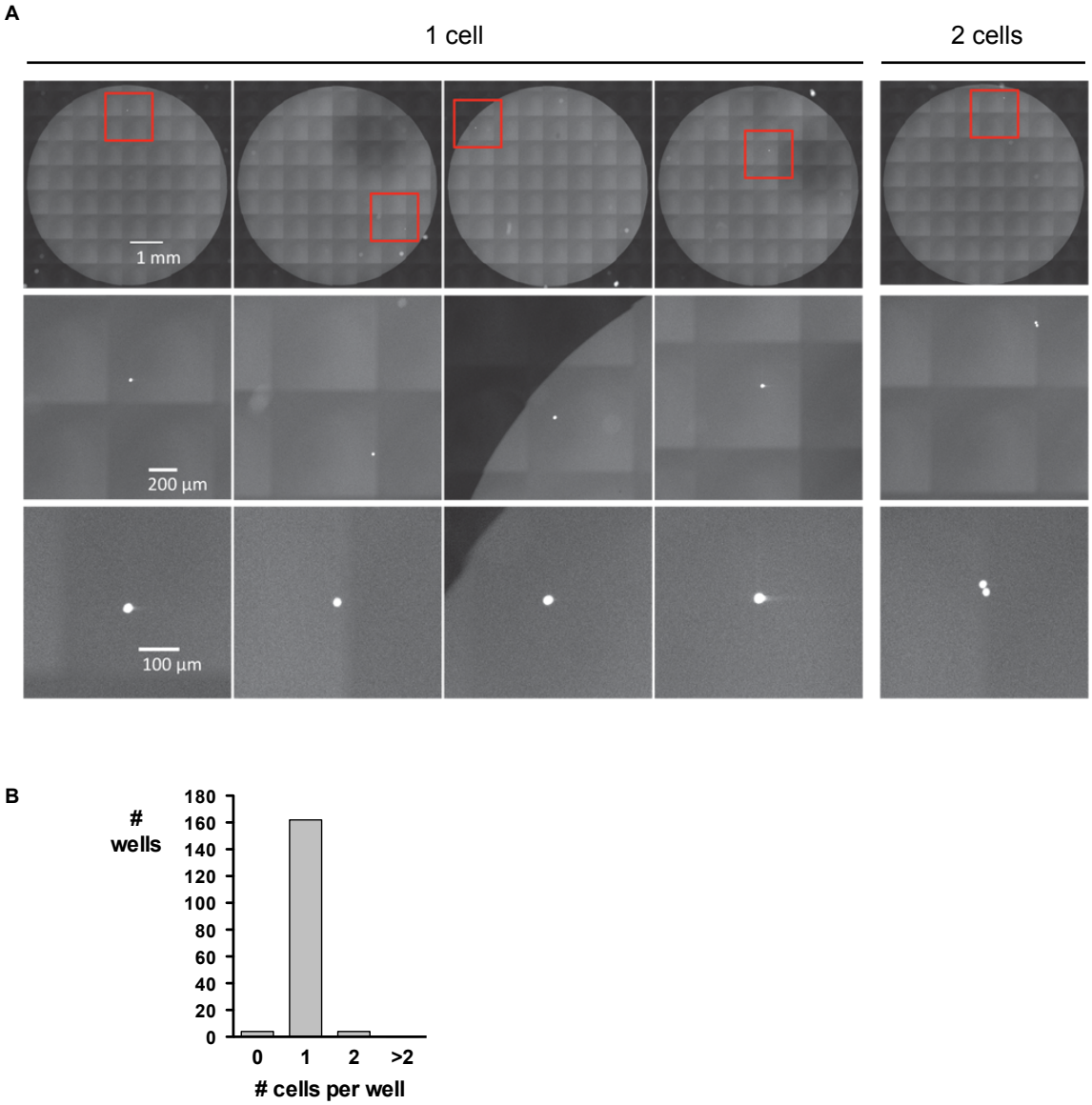


Figure S3. Single-cell sorting assessment. (A) Automated microscope scanning of CFSE-stained cells single cell-sorted into 96-well plates. Representative pictures and software-generated magnifications of wells containing one or two cells. **(B) Single-cell sorting quantification.** We sorted two 96-well plates in single-cell mode into all wells of columns 2 to 12 of each plate (column 1 was used for calibration; see Materials and Methods). The histogram shows the number of wells in which no cell (4 wells), one cell (162 wells), two cells (4 wells) or more than two cells (0) were detected (six wells were not properly scanned by the microscope and excluded from the analysis).

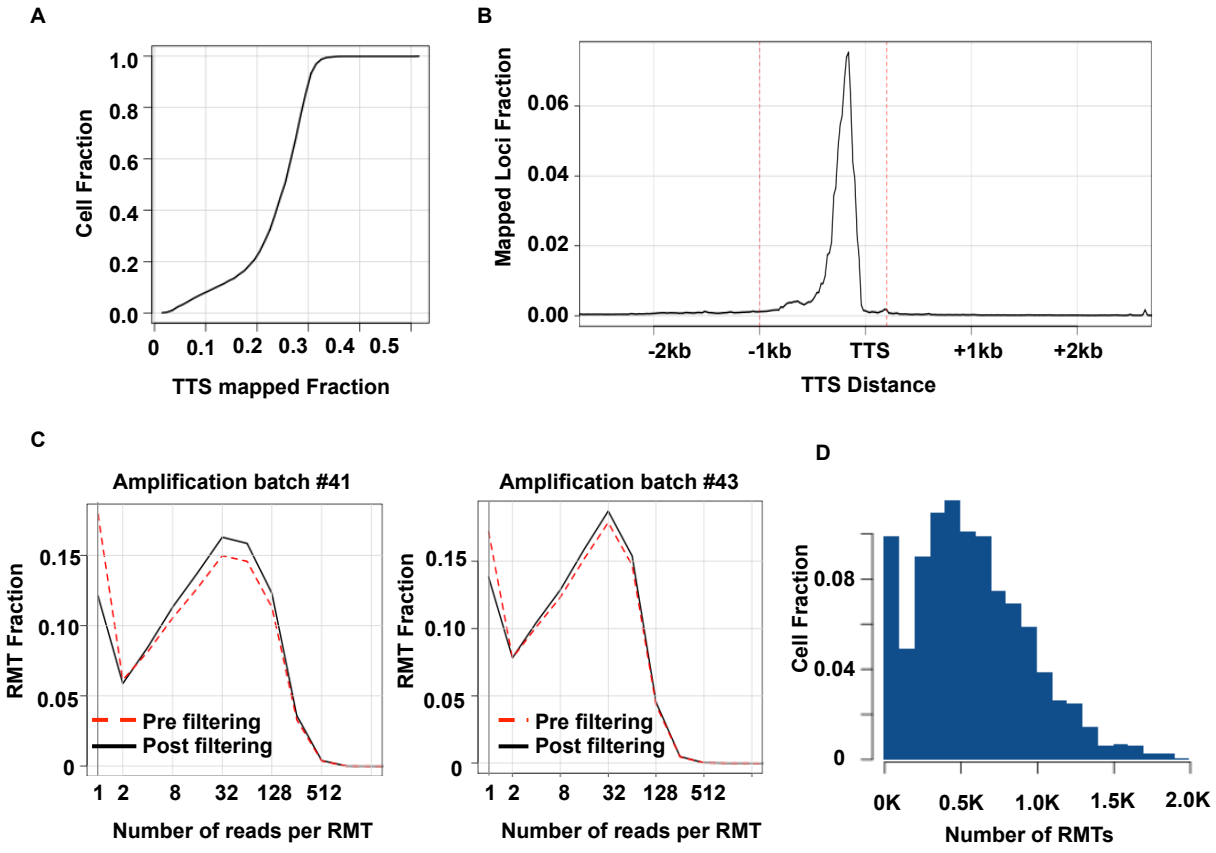


Figure S4. Sequence mapping and tag filtering. (A) TTS mapping efficiency. Cumulative distribution of the fraction of reads that were mapped to mouse TTS (x-axis, see methods) across 1528 multiplexed cells (y-axis show the cumulative cell fraction). Median mapping percentage is ~25%. **(B) Distribution of mapping loci around TTS.** Shown is the spatial distribution of mapped molecules (Cell/RMT) around TTS. Dashed red lines demarcate the (-1000, 200) range in which we consider a molecule as associated with a TTS. **(C) Read saturation.** Distribution of number of sequencing products per inferred molecule (unique and valid RMT) before (dashed red curve) and after (black solid curve) barcode and RMT filtering (see methods) for two amplification batches (i.e. batches sharing the same pool barcode). These data show that the majority of the molecules were excessively sequenced. **(D) RMT yield.** Shown are a distribution of RMTs after filtering (x-axis) and the fraction of cells (y-axis).

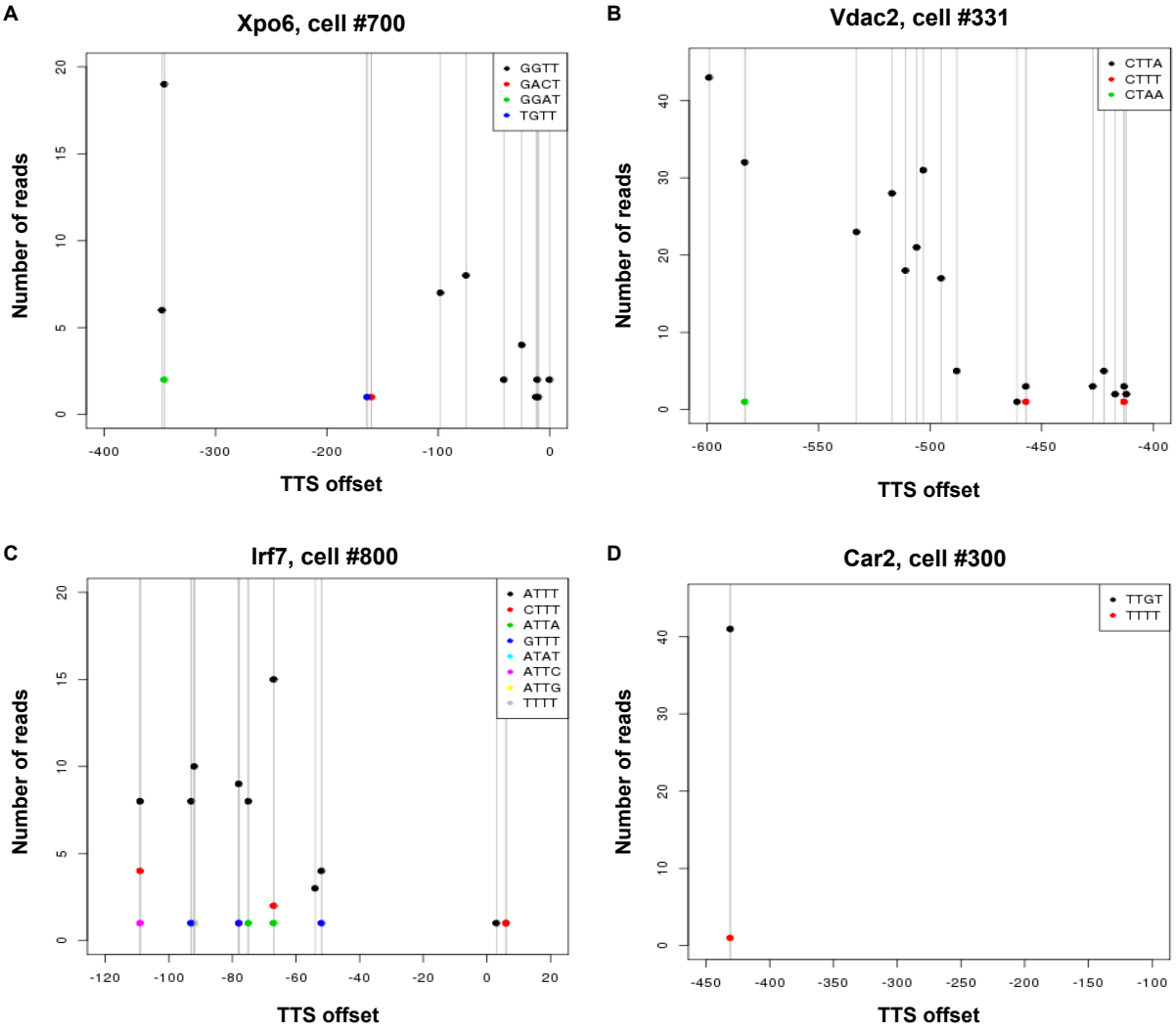


Figure S5. Filtering RMT sequencing errors. Shown are full sequencing and labeling data for 4 gene/cell examples. For each offset (x-axis) the number of reads of each color coded RMTs is shown. These profiles exemplify RMTs with multiple offsets (black dots, panels **A-C**) that undergo sequencing errors and create spurious RMT with edit distance of one (colored dots). Specifically in the Xpo6 example **(A)**, the GGTT RMT (black) is mapped to multiple positions with high read counts, but the GGAT is mapped to a single loci which is shared with GGTT. In some cases poorly mapped molecules undergo RMT sequencing errors (as shown in **(D)**). Detection and filtering of RMT sequence errors is described in the methods section.

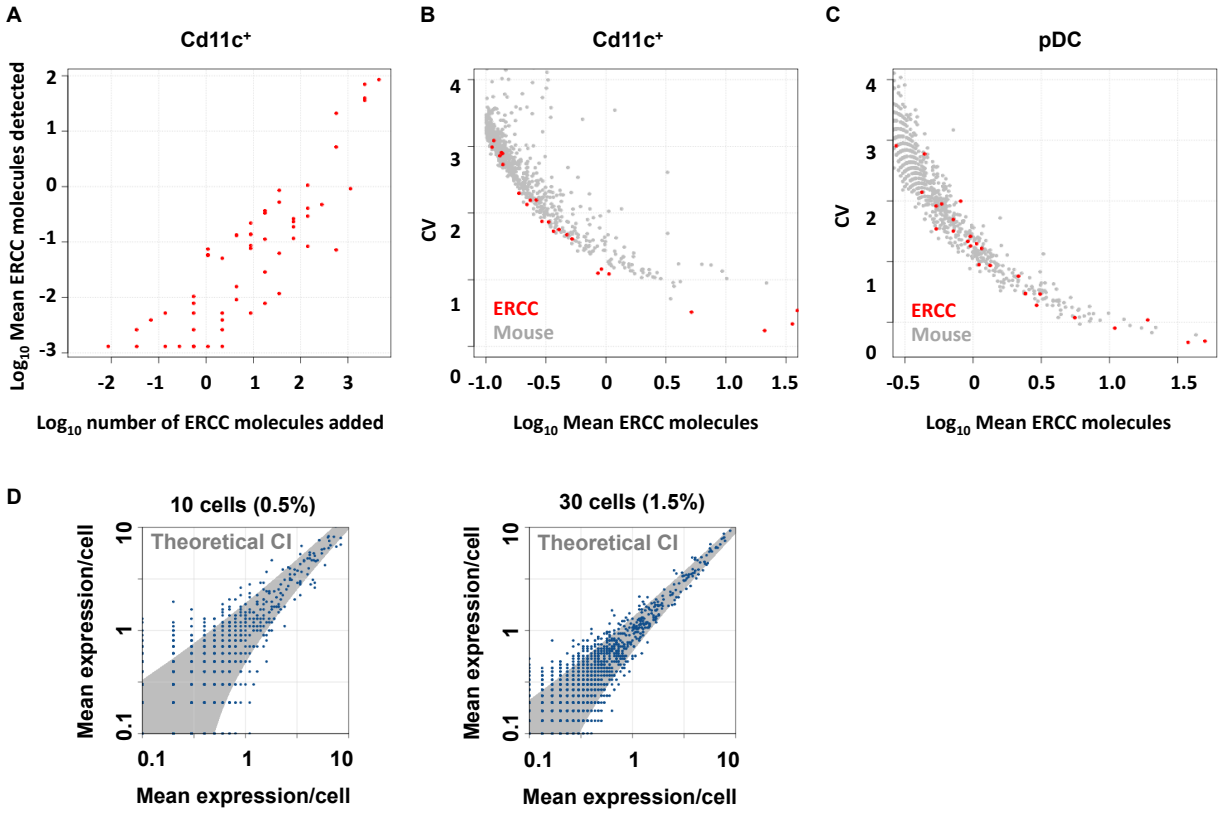


Figure S6. Technical variance. (A) ERCC RNA recovery of over four orders of magnitude. Shown is the average detection rate of ERCC spike-in molecules (y-axis) vs. the number of ERCC molecules added to each single cell (log scale) across 1536 CD11c$^+$ cells. The value for undetected molecule was set to (-3). The data reflect robust estimation of concentrations over ~4 orders of magnitudes, with some technical variability and provide bounds on the expected technical sequence specific recovery and sequencing bias in the protocol. Overall, following extensive filtering, we estimate recovery rate of 2-3% of the spiked-in molecules. **(B-C) Association between standard deviation and average of detected molecule counts.** Shown are coefficients of variance (CV, y-axis) vs. the average cellular mRNA (gray dots) and ERCC spike-in (red dots) molecule counts across 1536 Cd11c$^+$ cells (**B**) and 95 Cd8$^+$ pDCs (**C**). This analysis shows low technical variance between cellular mRNA and spike-in molecules compared to recently published methods (Wu et al. (*21*)). As expected, spike-in molecules have lower variability than cellular mRNA molecules, as their variance is only technical and does not involve a biological component. pDCs also show relatively homogeneous expression, especially compared to the heterogeneous CD11c+ dataset. Spike-in controls were processed using the same pipeline used for mouse sequences. (**D**) Similar to Fig 1C, but using 10, 30 cells.

We note that we recommend assessing technical variance in single-cell RNA-seq by plotting variance/mean against the mean, rather than the CV (as done in Fig. 1). This reflects the desire to see only technical binomial variance (that scale with the mean) in the data.

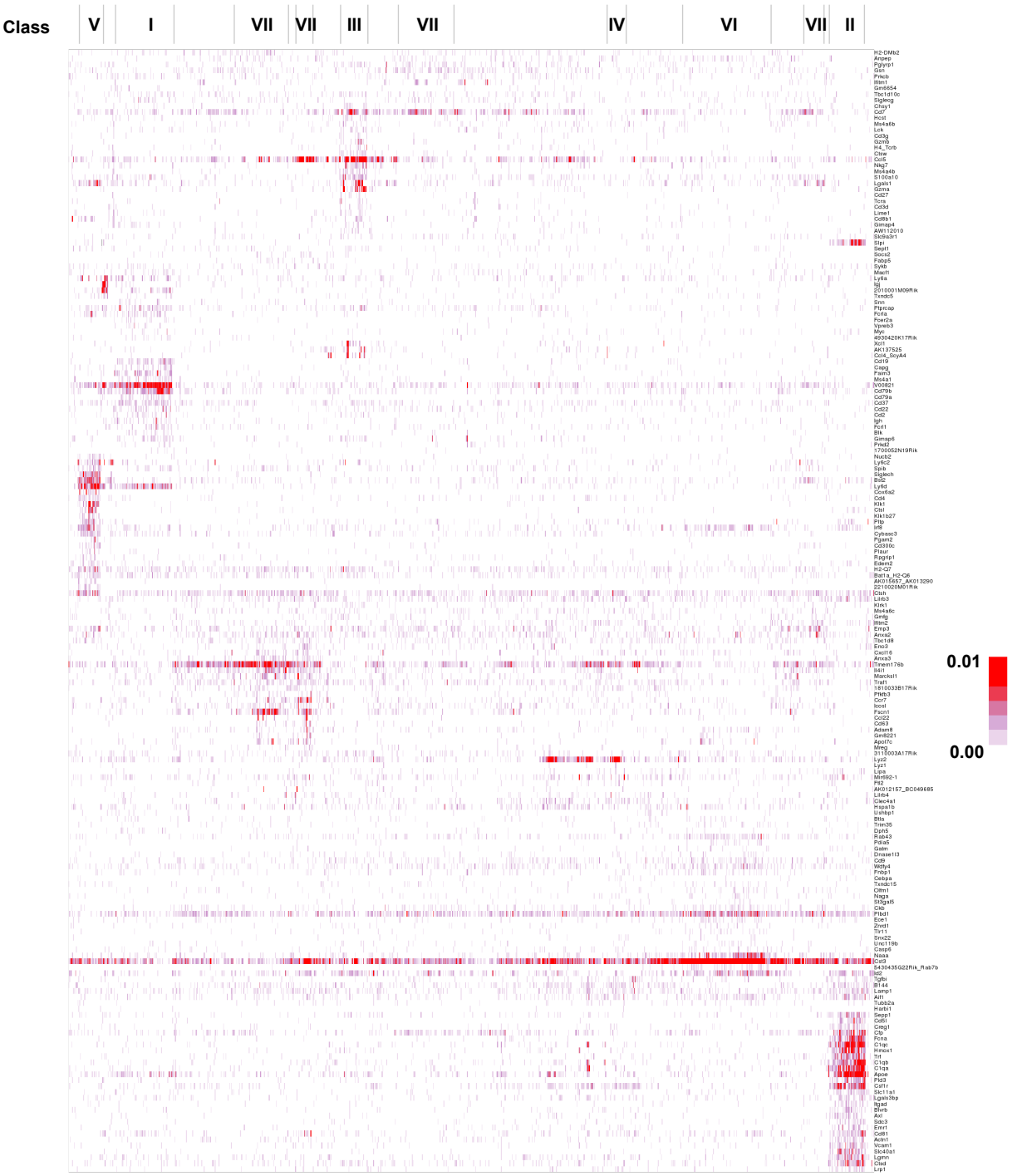


Figure S7. Gene-cell covariation structure over splenic cell population. Shown are color-coded (logarithmic scale) down-sampled molecular counts for selected genes (rows) over 1040 single cells (columns; ordering is identical to the clustered correlation map map of figure 2A). Groups of strongly correlated cells (as in figure 2A) are marked by black lines on top. This direct visualization of the dataset demonstrates how the correlation between cell type-specific gene-expression profiles combine to generate effective and usually unambiguous classification of cells into types. Once cell classes are identified, a much large number of cell-type specific genes are characterized through pooling together single cells.

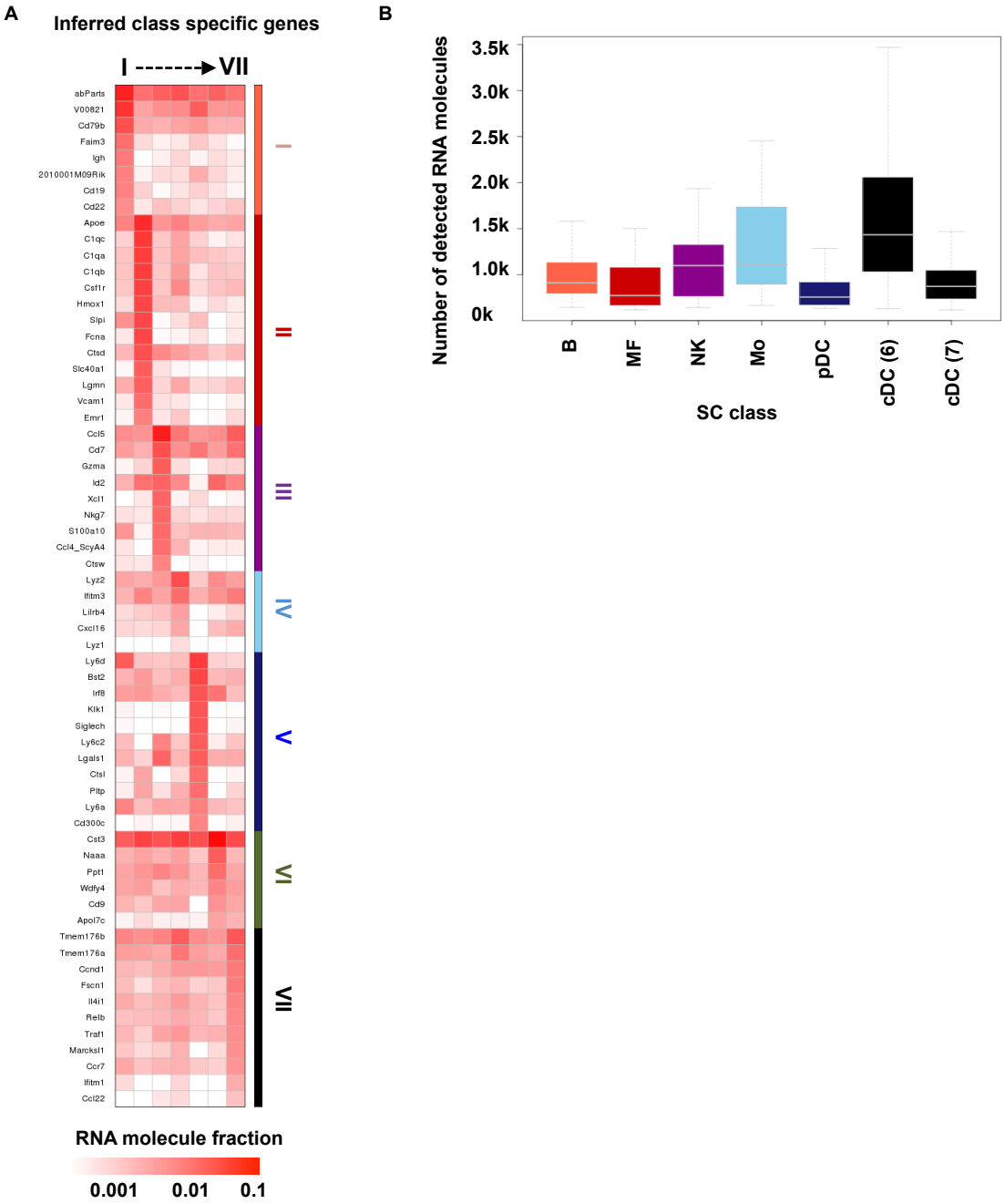


Figure S8. Mixture model for the CD11c+ enriched splenic cell population. (A) Selected cell-type specific genes for the spleen CD11c+ model. The color coded matrix depict class-specific mean expression for the CD11c^{+} model shown in Figure 2. Complete data for this model is provided in Table S2. **(B) Coverage varies among cell types.** Shown are the distributions pre-downsampled RMT coverage (Y axis) for over 2000 splenic cells, organized according to the inferred maximum *a posteriori* group (x-axis). While the model is inferred from uniform coverage data, different cell types are likely to provide markedly different numbers of recovered mRNAs.

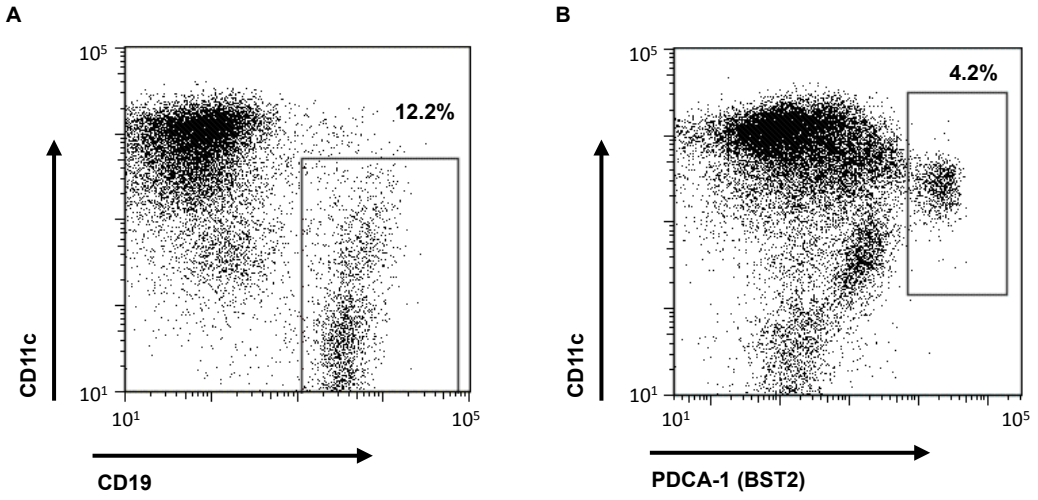


Figure S9. Non-DC frequency-estimation validation. FACS analysis was used to validate the estimated frequencies of B cells **(A)** and pDCs **(B)** in the $CD11c^+$ pool. Shown are independent experiments analyzing CD11c vs. CD19, a B cell marker, and PDCA-1 (Bst2), a pDC marker. B and pDC subpopulation frequencies are shown above the gating frame (gray).

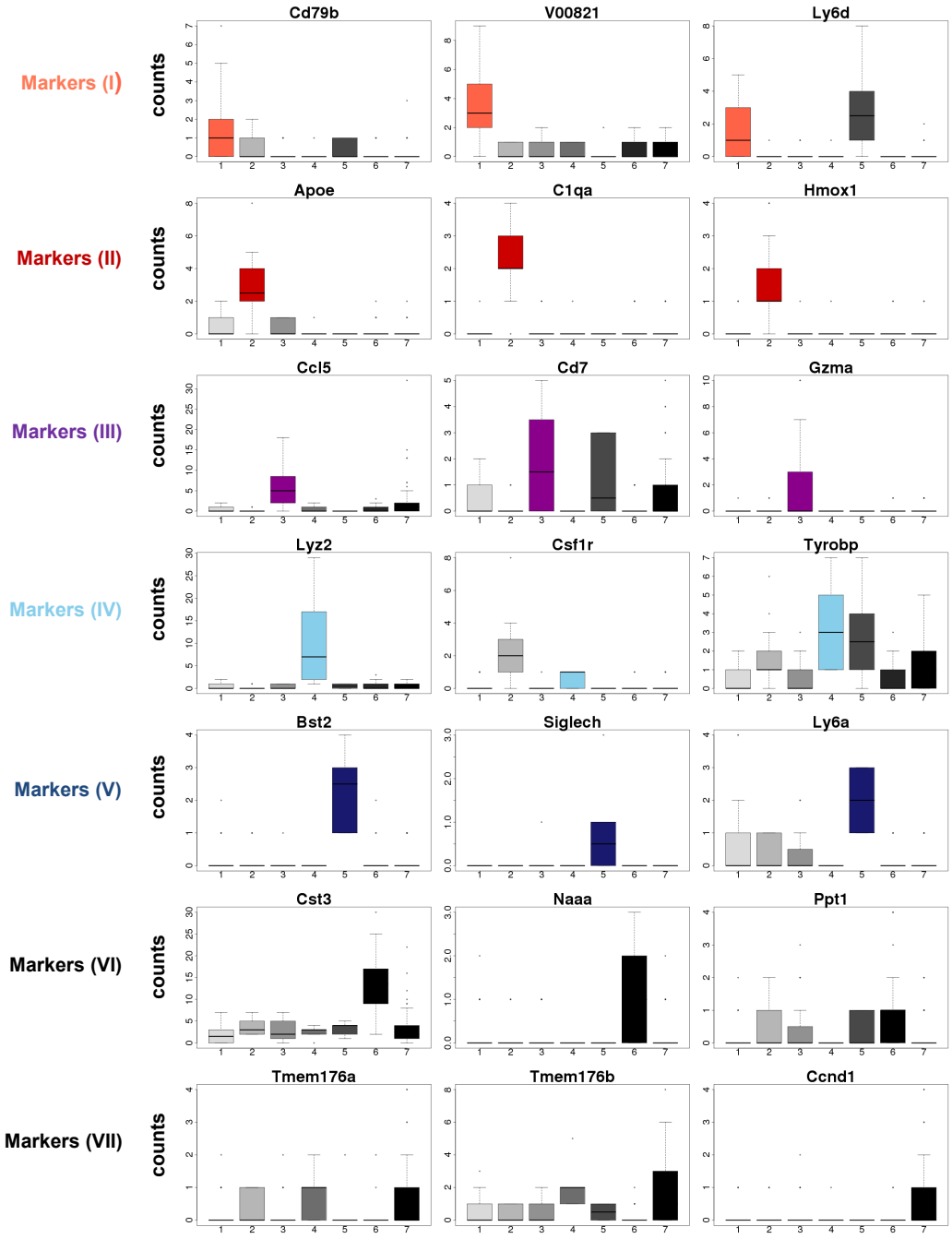


Figure S10. Expression distribution of representative genes shows variability between single-cell classes. Three strong "marker" genes were selected for each single-cell class. Shown here is their RMT coverage distribution in each of the seven classes (using down-sampled data to eliminate coverage difference between cell types).

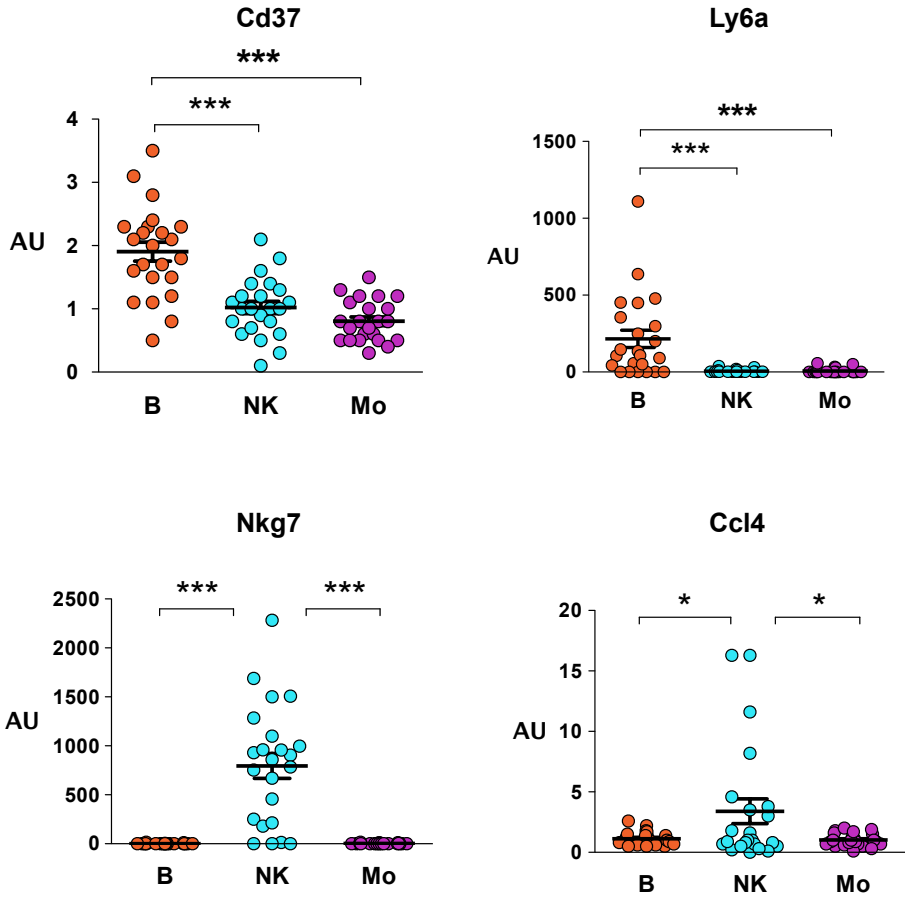


Figure S11. Single-cell gene expression validation by real-time PCR. Shown are four genes differentially expressed among three identified CD11c^{+} subpopulations (scatter plots) validated by single-cell RT pre-amplification real time qPCR. Error bars represent mean ± s.e.m. (n = 23 single cells); AU, arbitrary units; asterisks indicates ANOVA Bonferroni's Multiple Comparison test: *** P value < 0.001; * P value < 0.05

S12

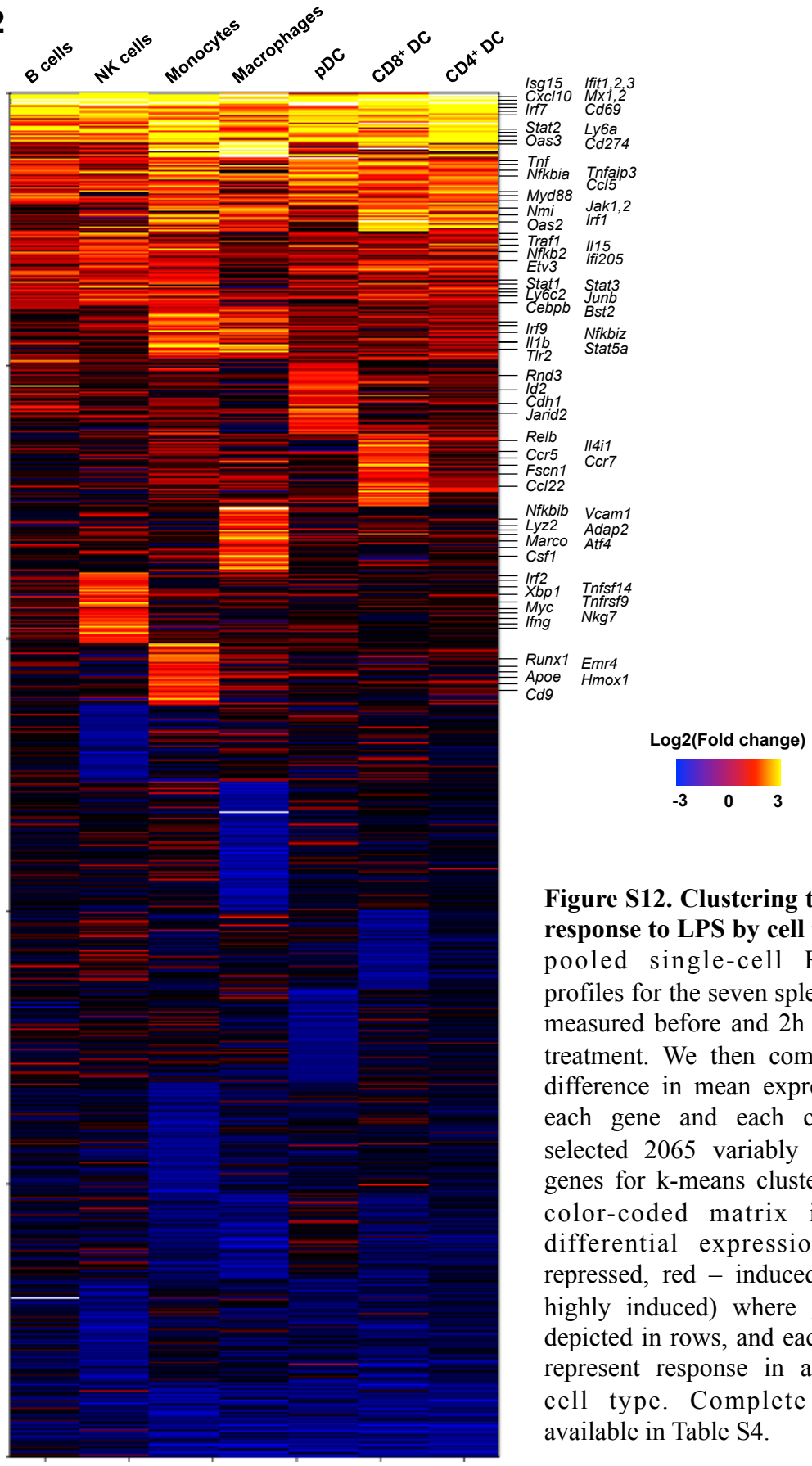


Figure S12. Clustering the spleen response to LPS by cell types. We pooled single-cell RNA-seq profiles for the seven splenic types, measured before and 2h after LPS treatment. We then computed the difference in mean expression for each gene and each class, and selected 2065 variably expressed genes for k-means clustering. The color-coded matrix indicates differential expression (blue-repressed, red – induced, yellow-highly induced) where genes are depicted in rows, and each column represent response in a different cell type. Complete data is available in Table S4.

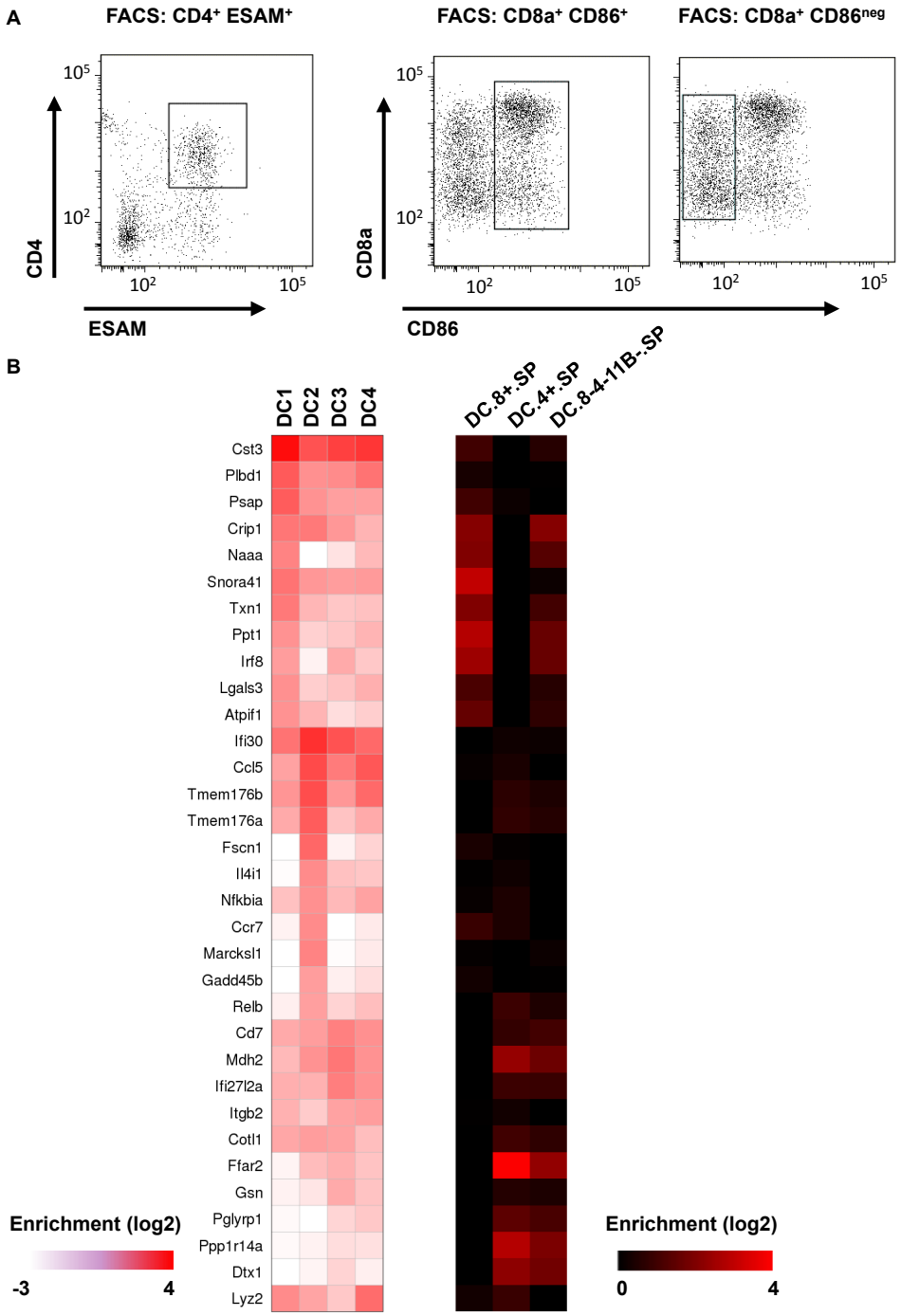


Figure S13. Marker-sorted DCs. (A) Comparison of FACS and single-cell RNA-based sorting was facilitated by FACS sorting and sequencing RNA from three DC subpopulations: CD8^{high} CD86^{+}, CD8^{inter} CD86^{-} and CD4^{+} ESAM^{+}. Gating is shown by gray boxes in the corresponding FACS plot. These sorted population were analyzed by single-cell RNA-seq and compared to the DC mixture model as shown in Fig 4B. **(B)** Shown are pooled single-cell mRNA mean counts (left) side by side with ImmGen gene expression data for three sorted DC classes (right). DC.8+.SP, splenic CD8a+ DCs; DC.4+.SP, splenic CD4+ DCs; DC.8-4-11B-.SP, splenic CD8a^{-} CD4^{-} CD11b^{-} (double negative) DCs. Genes that were specifically enriched in at least one of the three classes were selected for presentation. For the complete table of differentially expressed genes see table S5.

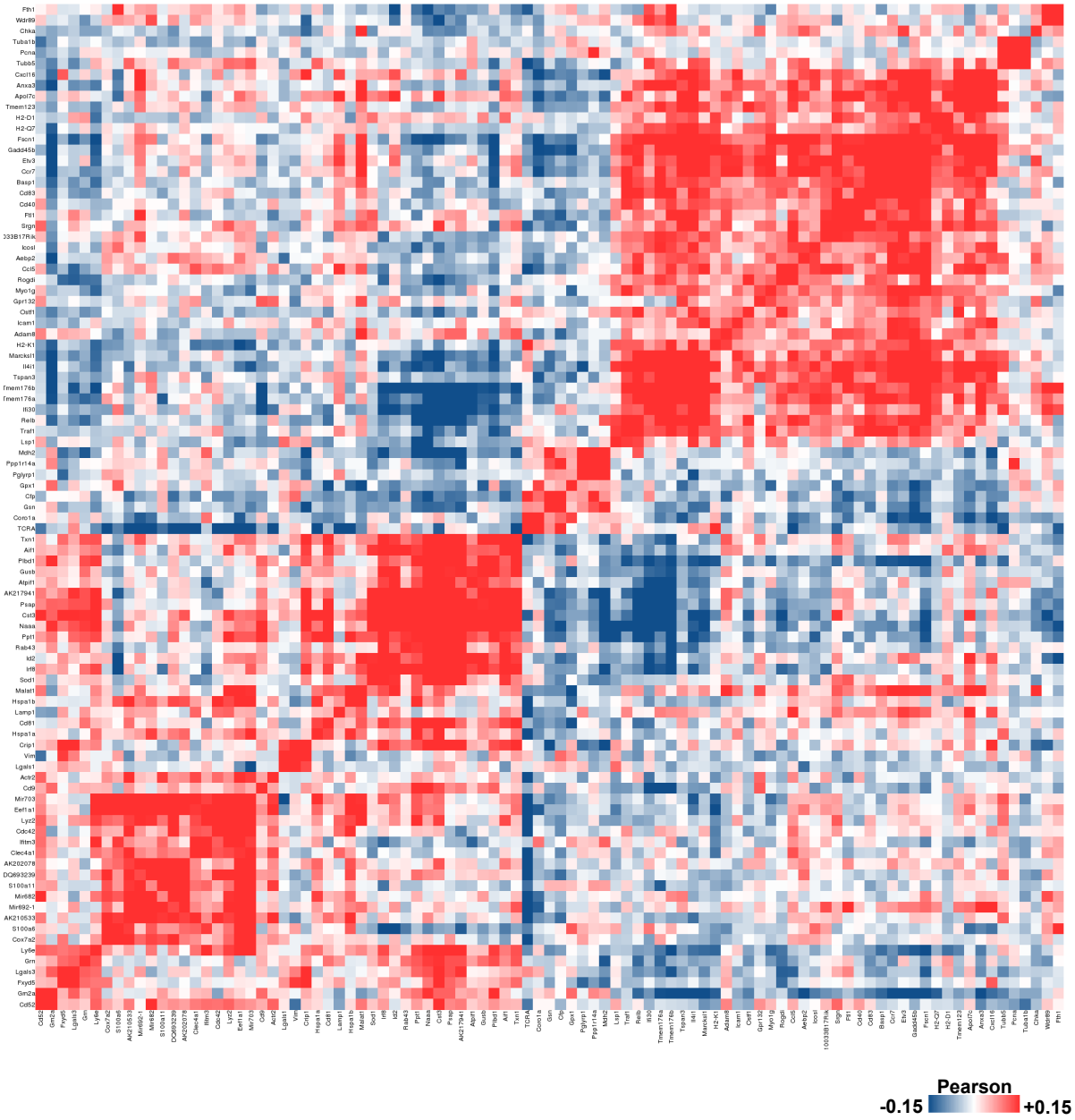


Figure S14. Shown is a gene correlation matrix depicting Pearson correlations between the single-cell RNA-seq profiles within 595 cells classified as DCs by our model. Only genes with at least 4 pairing with R > 0.14 are shown.

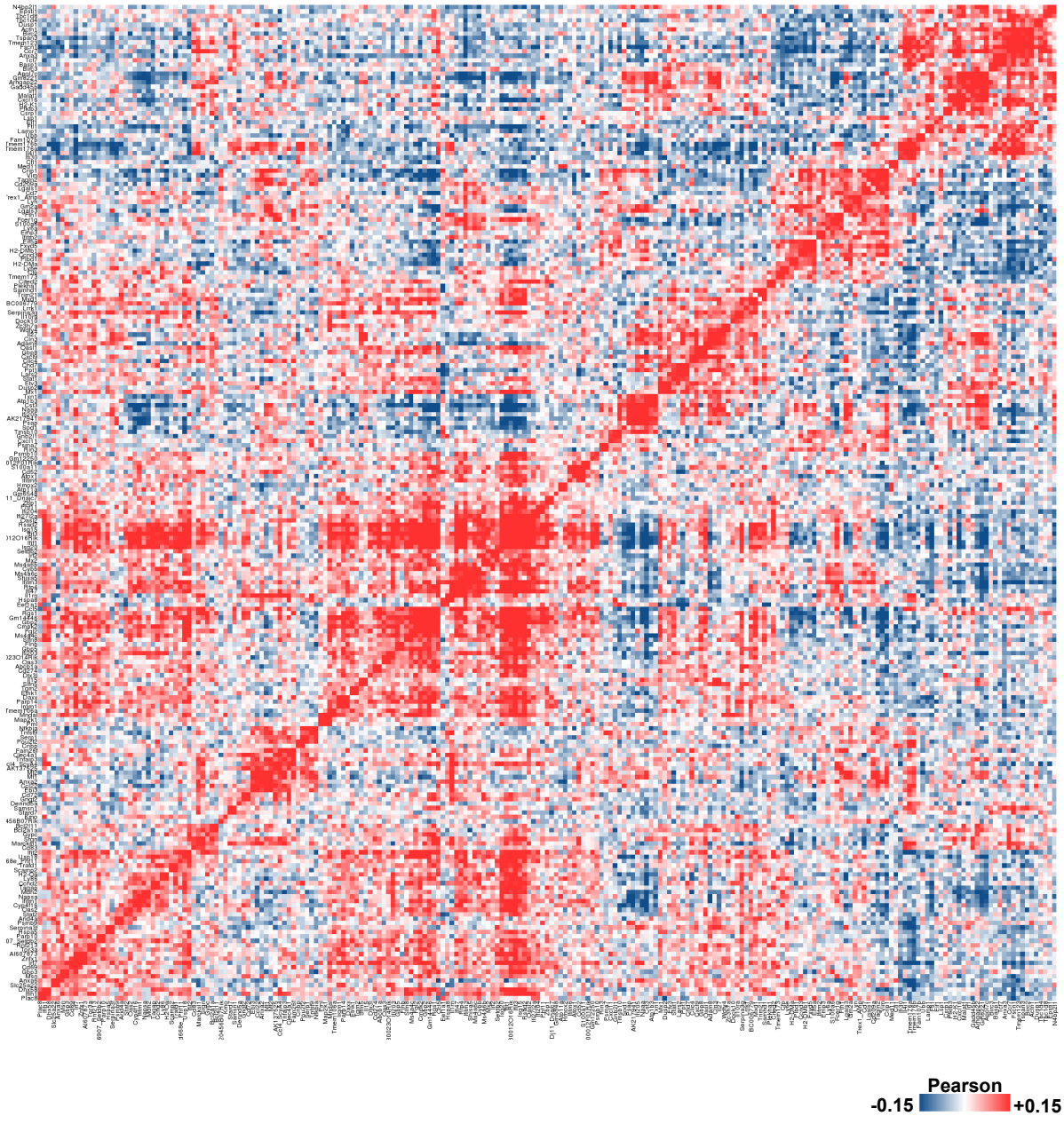


Figure S15. Shown is a gene correlation matrix depicting Pearson correlations between the *in vivo* LPS-treated single-cell RNA-seq profiles within 403 cells classified as DCs by our model. Only genes with at least 4 pairing with R > 0.15 are shown.

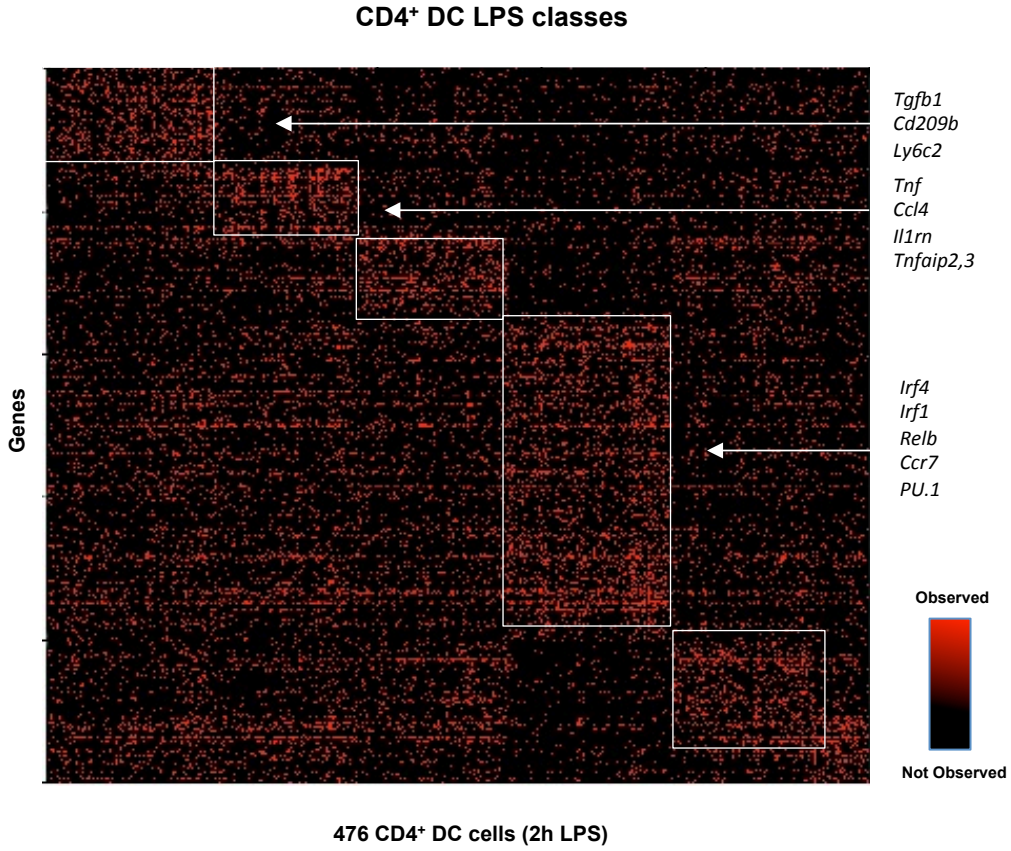


Figure S16. The heat map depicts mRNA counts in 476 single cells that were classified into class VII following *in vivo* LPS treatment. Cells were clusters using our EM-like iterative approach as described in the Methods. Genes are grouped according to the class in which they are mostly enriched. Complete data is available in Table S6.

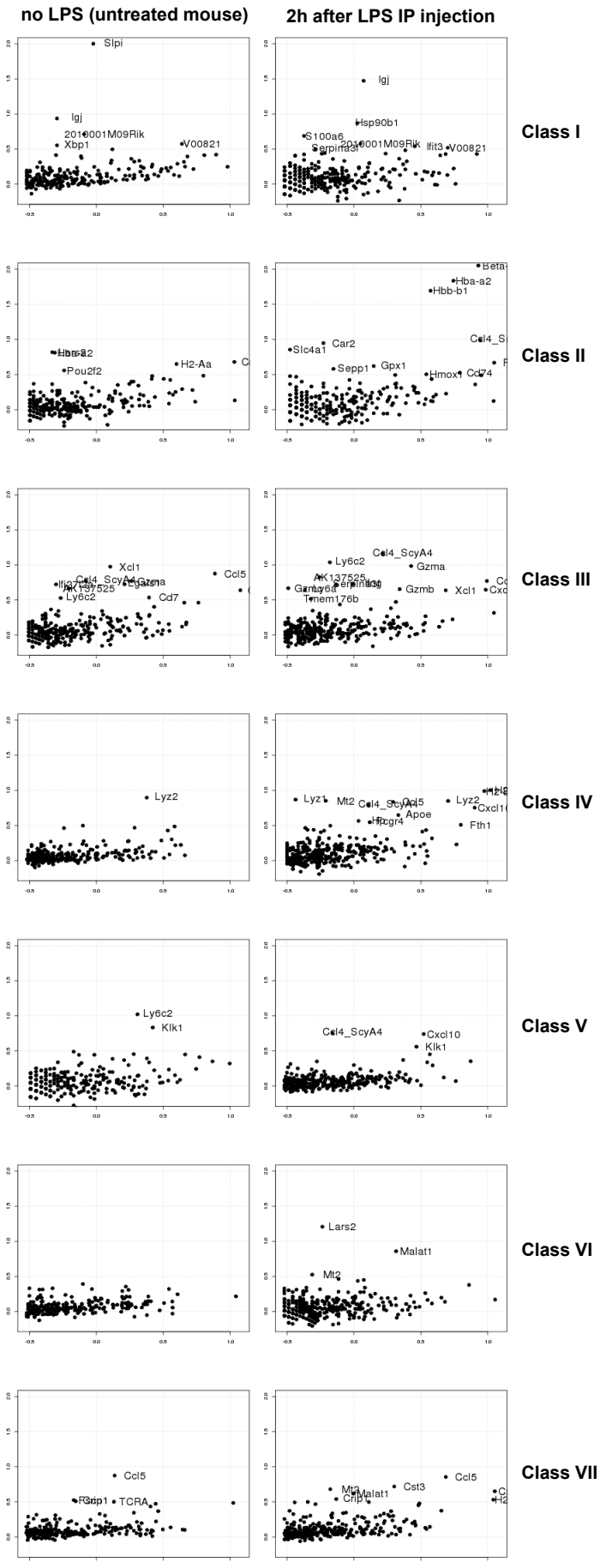


Figure S17. For each of the seven splenic cell classes, we show the mean and variance of all genes, before and after LPS treatment. Genes showing high variance indicate potential further organization of these populations into subtypes. Genes affected following LPS treatment indicate potential re-organization of the cell population into specific response classes, as outlined above in more detail for $CD4^{+}$ DCs (Fig S16).

Supplementary Tables

Table S7. Primers used during MARS-seq library construction

Primer name	Sequence and modifications
barcoded RT primer	CGATTGAGGCCGGTAATACGACTCACTATAGGGGCGACGTGT GCTCTTCCGATCTXXXXXXNNNNTTTTTTTTTTTTTTTTTTN, where XXXXXX is the cell barcode and NNNN is the RMT
ligation adapter	AGATCGGAAGAGCGTCGTGTAG, modified with a phosphate group at 5' and a C3 spacer (blocker) at the 3'
Second RT primer	TCTAGCCTTCTCGCAGCACATC
P5_Rd1 PCR forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC GACGCTCTTCCGATCT
P7_Rd2 PCR reverse	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGAC GTGTGCTCTTCCGATCT

Table S8. Indexed ligation adapters for 384-well plates

Primer name	Sequence and modifications
lig_NNNX4_ix1	/5Phos/GACTNNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix2	/5Phos/CATGNNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix3	/5Phos/CCAANNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix4	/5Phos/CTGTNNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix5	/5Phos/GTAGNNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix6	/5Phos/TGATNNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix7	/5Phos/ATCANNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/
lig_NNNX4_ix8	/5Phos/TAGANNNAGATCGGAAGAGCGTCGTGTAG/3SpC3/

Table S9. RT-preamplification primers

Primer name	Sequence
CD37-F	CTGTCTCCTGGGCCTGTATT
CD37-R	CACCAATTCCTGCACCCTTC
NKg7-F	GTTCTGTCTTGCATCCCAGC
NKg7-R	CTGGCTCCATCTCATACTGGT
Ly6A-F	AATTACCTGCCCCTACCCTG
Ly6A-R	GCAGATGGGTAAGCAAAGATTG
Ccl4-F	TGTGCTCCAGGGTTCTCAG
Ccl4-R	AATCCATCACAAAGCTTCTGTG

References and Notes

1. M. Acar, J. T. Mettetal, A. van Oudenaarden, Stochastic switching as a survival strategy in fluctuating environments. *Nat. Genet.* **40**, 471–475 (2008). doi:10.1038/ng.110 Medline

2. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002). doi:10.1126/science.1070919

3. R. N. Germain, Maintaining system homeostasis: The third law of Newtonian immunology. *Nat. Immunol.* **13**, 902–906 (2012). doi:10.1038/ni.2404 Medline

4. S. C. Bendall, E. F. Simonds, P. Qiu, A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe'er, S. D. Tanner, G. P. Nolan, Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011). doi:10.1126/science.1198704

5. B. M. Bradford, D. P. Sester, D. A. Hume, N. A. Mabbott, Defining the anatomical localisation of subsets of the murine mononuclear phagocyte system using integrin alpha X (Itgax, CD11c) and colony stimulating factor 1 receptor (Csf1r, CD115) expression fails to discriminate dendritic cells from macrophages. *Immunobiology* **216**, 1228–1237 (2011). doi:10.1016/j.imbio.2011.08.006 Medline

6. F. Geissmann, S. Gordon, D. A. Hume, A. M. Mowat, G. J. Randolph, Unravelling mononuclear phagocyte heterogeneity. *Nat. Rev. Immunol.* **10**, 453–460 (2010). doi:10.1038/nri2784 Medline

7. D. A. Hume, Applications of myeloid-specific promoters in transgenic mice support in vivo imaging and functional genomics but do not support the concept of distinct macrophage and dendritic cell lineages or roles in immunity. *J. Leukoc. Biol.* **89**, 525–538 (2011). doi:10.1189/jlb.0810472 Medline

8. M. C. Nussenzweig, R. M. Steinman, M. D. Witmer, B. Gutchinov, A monoclonal antibody specific for mouse dendritic cells. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 161–165 (1982). doi:10.1073/pnas.79.1.161 Medline

9. R. M. Steinman, Z. A. Cohn, Identification of a novel cell type in peripheral lymphoid organs of mice. I. Morphology, quantitation, tissue distribution. *J. Exp. Med.* **137**, 1142–1162 (1973). doi:10.1084/jem.137.5.1142 Medline

10. L. Bar-On, T. Birnberg, K. L. Lewis, B. T. Edelson, D. Bruder, K. Hildner, J. Buer, K. M. Murphy, B. Reizis, S. Jung, $CX3CR1^{+}$ $CD8\alpha^{+}$ dendritic cells are a steady-state population related to plasmacytoid dendritic cells. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14745–14750 (2010). doi:10.1073/pnas.1001562107 Medline

11. D. Hashimoto, J. Miller, M. Merad, Dendritic cell and macrophage heterogeneity in vivo. *Immunity* **35**, 323–335 (2011). doi:10.1016/j.immuni.2011.09.007 Medline

12. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012). doi:10.1016/j.celrep.2012.08.003 Medline

13. S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J. B. Fan, P. Lönnerberg, S. Linnarsson, Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828 (2012). doi:10.1038/nprot.2012.022 Medline

14. D. Ramsköld, S. Luo, Y. C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, R. Sandberg, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012). doi:10.1038/nbt.2282 Medline

15. Y. Sasagawa, I. Nikaido, T. Hayashi, H. Danno, K. D. Uno, T. Imai, H. R. Ueda, Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, R31 (2013). doi:10.1186/gb-2013-14-4-r31 Medline

16. A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, A. Regev, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013). doi:10.1038/nature12172 Medline

17. F. Tang, K. Lao, M. A. Surani, Development and applications of single-cell transcriptome analysis. *Nat. Methods* **8**, (Suppl), S6–S11 (2011). Medline

18. A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke, S. R. Quake, Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014). doi:10.1038/nmeth.2694 Medline

19. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014). doi:10.1126/science.1245316

20. S. Islam *et al.*, *Nat. Methods*, published online 22 December 2013 (10.1038/nmeth.2772).

21. Materials and methods are available as supplementary materials on *Science* Online.

22. I. Amit, M. Garber, N. Chevrier, A. P. Leite, Y. Donner, T. Eisenhaure, M. Guttman, J. K. Grenier, W. Li, O. Zuk, L. A. Schubert, B. Birditt, T. Shay, A. Goren, X. Zhang, Z. Smith, R. Deering, R. C. McDonald, M. Cabili, B. E. Bernstein, J. L. Rinn, A. Meissner, D. E. Root, N. Hacohen, A. Regev, Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–263 (2009). doi:10.1126/science.1179050

23. T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012). doi:10.1038/nmeth.1778 Medline

24. S. C. Baker, S. R. Bauer, R. P. Beyer, J. D. Brenton, B. Bromley, J. Burrill, H. Causton, M. P. Conley, R. Elespuru, M. Fero, C. Foy, J. Fuscoe, X. Gao, D. L. Gerhold, P. Gilles, F. Goodsaid, X. Guo, J. Hackett, R. D. Hockett, P. Ikonomi, R. A. Irizarry, E. S. Kawasaki, T. Kaysser-Kranich, K. Kerr, G. Kiser, W. H. Koch, K. Y. Lee, C. Liu, Z. L. Liu, A. Lucas, C. F. Manohar, G. Miyada, Z. Modrusan, H. Parkes, R. K. Puri, L. Reid, T. B. Ryder, M. Salit, R. R. Samaha, U. Scherf, T. J. Sendera, R. A. Setterquist, L. Shi, R.

Shippy, J. V. Soriano, E. A. Wagar, J. A. Warrington, M. Williams, F. Wilmer, M. Wilson, P. K. Wolber, X. Wu, R. Zadro, External RNA Controls Consortium, a progress report. *Nat. Methods* **2**, 731–734 (2005). doi:10.1038/nmeth1005-731 Medline

25. P. Dalerba, T. Kalisky, D. Sahoo, P. S. Rajendran, M. E. Rothenberg, A. A. Leyrat, S. Sim, J. Okamoto, D. M. Johnston, D. Qian, M. Zabala, J. Bueno, N. F. Neff, J. Wang, A. A. Shelton, B. Visser, S. Hisamori, Y. Shimono, M. van de Wetering, H. Clevers, M. F. Clarke, S. R. Quake, Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011). doi:10.1038/nbt.2038 Medline