nature biotechnology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41587-020-0465-8

In the format provided by the authors and unedited.

Systematic comparison of single-cell and single-nucleus RNA-sequencing methods

Jiarui Ding[1], Xian Adiconis[1,9], Sean K. Simmons[1,9], Monika S. Kowalczyk[1], Cynthia C. Hession[1], Nemanja D. Marjanovic[1], Travis K. Hughes[1,2,3,4], Marc H. Wadsworth[1,2,3,4], Tyler Burks[1], Lan T. Nguyen[1], John Y. H. Kwon[1], Boaz Barak[5], William Ge[1], Amanda J. Kedaigle[1], Shaina Carroll[1,2,3,4], Shuqiang Li[1], Nir Hacohen[1,6], Orit Rozenblatt-Rosen[1], Alex K. Shalek[1,2,3,4], Alexandra-Chloé Villani[1,6,7], Aviv Regev[1,4,8] and Joshua Z. Levin[1]✉

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Institute for Medical Engineering and Science, Department of Chemistry, MIT, Cambridge, MA, USA. [3]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. [4]Koch Institute of Integrative Cancer Research, Cambridge, MA, USA. [5]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA. [6]Center for Cancer Research, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [7]Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, MA, USA. [8]Howard Hughes Medical Institute, Department of Biology, MIT, Cambridge, MA, USA. [9]These authors contributed equally: Xian Adiconis, Sean K. Simmons. ✉e-mail: jlevin@broadinstitute.org

Supplementary Note 1

Assessing presence of poly(T) in reads

The methods varied in the fraction of reads without poly(T) at the expected positions. Methods that do not use beads to capture mRNA (CEL-Seq2 and sci-RNA-seq) generally had poly(T) detected at the expected positions (**Extended Data Fig. 3**, **Supplementary Tables 3** and **4**). The 10x Chromium data also had a high fraction of reads with poly(T) at the expected position. By contrast, Drop-seq, Seq-Well, and inDrops showed high percentages of reads without poly(T). Interestingly, although Drop-seq and Seq-Well used the same batch of beads, they differ in the percentages of reads without poly(T): 6.5% and 7.5% for Drop-seq, but 10.8% and 16.8% for Seq-Well in the mixture experiments. Similarly, in the PBMC experiments, 10x Chromium (v2 and v3) had the lowest fraction of reads lacking poly(T) (0.7% to 2.1%) and Seq-Well had the highest (41.2%; **Supplementary Table 4**, **Extended Data Fig. 3b**). These results suggest that some aspect of the methods independent of the beads contributed to the proportion of reads without poly(T). In the nuclei experiments, DroNc-seq had the highest fraction of reads lacking poly(T) with nuclei (**Supplementary Tables 3-5**, **Extended Data Fig. 3c**).

Supplementary Note 2

Investigating antisense reads

Interestingly, 10x Chromium (v2) had the highest fraction of antisense reads (33% and 29%), followed by DroNc-seq (10% and 12%), and sci-RNA-seq (6% and 9%; **Extended Data Fig. 3c**). We did not observe this level of antisense reads in bulk cortex nuclei nor in PBMC scRNA-seq datasets (data not shown). To explore further, we analyzed all reads that aligned in an antisense orientation to transcripts that did not overlap any other transcripts (on either strand) and measured the fraction of reads with the sequence AAAAA at a distance of 1 to 500 bp upstream

from the start. We found that 10x Chromium, but not DroNc-seq or sci-RNA-seq, had a large peak in density of AAAAA sequences ~200bp away (**Supplementary Fig. 8**), suggesting these may result from internal priming at poly(A) sequences, perhaps from the first stand cDNA. We excluded these antisense reads from our subsequent analyses.

Supplementary Note 3

Fraction of mitochondrial reads varies, but is at or below level in bulk RNA-seq

We determined the fraction of reads aligning to mitochondrial genes, as these can be associated with lower quality scRNA-seq samples[1] in some cases, reflecting cellular stress, but can also be useful for inferring lineage relations between cells based on mtDNA mutations[2]. We reasoned that the mitochondrial ratios from the bulk RNA-seq could be a 'ground truth,' reflecting the actual mitochondrial transcript ratios in cells given the properties of the cells under consideration, without extra steps unique to scRNA-seq (be it dissociation, nucleus isolation, cell sorting, cell encapsulation or specific molecular biology steps).

In the mixture experiments, CEL-Seq2 had the highest frequency of such reads at 6.8 and 7.2% (**Supplementary Fig. 6a**,**b**), but this is similar to the levels of mitochondrial transcripts in bulk RNA-seq for the cell line samples: 13.0% and 15.3% for NIH3T3 cells, and 8.5% and 9.1% for HEK293 cells, suggesting this may be an accurate reflection of the cells' transcriptome. In PBMCs, inDrops (9.9% and 7.9%) and CEL-Seq2 (11.9% and 12.9%) had the highest mitochondrial ratios (**Supplementary Fig. 6b**). This may reflect the fact that both methods use *in vitro* transcription amplification rather than PCR. In addition, 10x Chromium (v3) had a high mitochondrial ratio at 11.2% (**Supplementary Fig. 6b**). Bulk RNA-seq of PBMC1 and PBMC2 had 12.6% and 12.3% mitochondrial transcripts, respectively, again suggesting this may be an accurate reflection of the transcriptome. In cortex nuclei, all methods had low mitochondrial

ratios (**Supplementary Fig. 6c**), consistent with their low (2.2% and 1.9%) fraction in matched bulk RNA-seq samples, and likely reflecting the removal of most mitochondria during nucleus isolation[3]. It remains to be proven whether a mitochondrial ratio more similar to bulk samples is a better reflection of the actual transcriptome of the cells in a given sample.

Supplementary Note 4

Comparison of published datasets

To assess whether our results generalize to other datasets, we analyzed published datasets in which two of these methods were tested on the same tissue (**Supplementary Table 12**). In pancreas tissue, CEL-Seq2[4] had somewhat higher sensitivity than Smart-seq2[5] (**Supplementary Fig. 7**). In placental villi and decidua, 10x Chromium had greater sensitivity than Drop-seq (**Supplementary Fig. 7**), as reported[6]. We note however that this analysis has caveats, including differences in the scRNA-seq protocol between ours and the published study, e.g., Smart-seq2, and that the samples in the published studies were not processed in a deliberately controlled manner for comparison purposes.

Supplementary Note 5

Reproducibility and accuracy

We compared the reproducibility of expression quantifications from replicates, among the different methods. In the mixture experiments, we calculated the Pearson correlation coefficients between pseudo-bulk data of each of the scRNA-seq mixture human and mouse datasets (**Online Methods**). As expected, replicates were almost always more correlated with each other than with pseudo-bulks from different methods (**Supplementary Fig. 4**). The pseudo-bulks from sci-RNA-seq, although correlated well with each other from the two replicates, were not as correlated with

the pseudo-bulks from other methods. The pseudo-bulks from the other methods were highly correlated with each other (**Supplementary Fig. 4**), suggesting that all methods produce highly consistent results in quantifying gene expression, as seen in previous comparison studies that demonstrated scRNA-seq methods had high accuracies based on measured ERCC spike-ins and their annotated molarities[7, 8]. Notably, pairs of methods with similar molecular biology – CEL-Seq2 with inDrops and Drop-seq with Seq-Well – were more highly correlated with each other (**Supplementary Fig. 4**).

To further explore reproducibility, we looked for genes differentially expressed between the two experiments for each scRNA-seq method (sampling the same average number of UMIs per cell, **Online Methods**). We did this separately for the human and mouse cells and we identified many differentially expressed genes for each method (**Supplementary Tables 13** and **14**). In particular, Mixture1 has much higher expression of immediate early genes (e.g., *Fosb*, *Fos*, *Egr1*) than Mixture2 for most methods, though not Smart-seq2 and only slightly higher for Seq-Well (**Supplementary Fig. 9**). This suggests that Mixture1 cells may have undergone more stress, possibly during cell isolation.

Differential expression analysis among the scRNA-seq methods in the mixture datasets (for human and mouse separately) showed GC content and length differences of these differentially expressed genes among the methods. This suggests potential biases for the methods (**Supplementary Fig. 10)** and is consistent with previous findings[9]. We note that the mixture datasets are the best suited in our study for this analysis because they are not confounded by having different proportions of cell types.

To examine how well the scRNA-seq methods reflect gene expression captured in bulk experiments, we compared them in the PBMC datasets. For each library, we combined each of

the single cell datasets, sampled to the same number of reads per cell, into a pseudo-bulk dataset and compared each of them to the actual matched bulk (**Online Methods**). Most libraries showed a strong correlation with the bulk, with Smart-seq2, 10x Chromium, and inDrops having the highest correlations (**Supplementary Fig. 11**). In addition, to assess cell-to-cell variation in gene expression, we determined the distribution of correlation of gene expression for each cell in the mixture experiments with the bulk data (**Supplementary Fig. 12**). In general, this metric appears to reflect the relative sensitivity of the scRNA-seq method (**Fig. 2b**).

Supplementary Note 6

Pooled data analysis across methods for mouse cortex nuclei

The combined analysis of mouse cortex nuclei (**Supplementary Fig. 5a**) allowed us to identify all the cell types in each dataset, except for pericytes, which clustered with endothelial cells (**Supplementary Fig. 5b**,**c**). In most cases, the combined and individual assignments agree (**Fig. 6**, **Supplementary Fig. 5c**). The 10x Chromium (v2) was the most consistent between the joint and individual level clustering, while sci-RNA-seq was the least (**Supplementary Fig. 5c**). In particular, some cell types were harder to distinguish—for example, oligodendrocytes and oligodendrocyte precursors, and microglia and excitatory neurons. All the cell clusters assigned new cell types by Harmony were confirmed when cells from individual libraries were checked with our AUC method (**Online Methods**).

Supplementary Note 7

Comparison of scumi with standard computational pipelines

There are significant differences in how the individual method-specific pipelines work. For example, the Drop-seq pipeline[10] only considers primary alignments and discards secondary

ones. For the reads that can be equally mapped to multiple positions, the primary alignment is typically randomly chosen from these alignments. On the other hand, Cell Ranger[11] (for 10x Chromium data) only counts those multi-mapped reads if all the alignments of a read overlap a single gene.

Generally, there was good agreement between the basic metrics generated by the method-specific and scumi pipelines in terms of the median number of UMIs and genes per cell for the PBMC datasets (**Supplementary Table 4**). In particular, the relative ranking of all the methods for median genes per cell is the same between the scumi and method-specific computational pipelines (**Supplementary Table 4**). The most substantial discrepancy between the standard and scumi pipelines is for inDrops, where the standard pipeline produces much higher estimates of the number of UMIs and genes. This, however, is to be expected as the inDrops pipeline goes to great lengths to recover reads that might be discarded otherwise. For example, for reads that map to more than one gene, the inDrops pipeline uses information about the distances to the 3' end of genes to choose one mapping as optimal, while scumi does not. One of the most challenging steps in the analysis is choosing the correct number of cells in a dataset and most of these method-specific pipelines do not include this as a formal part of their pipeline or it does not work that well in practice, so that we followed standard practices and manually chose the number of cells based on the number of genes per cell in most cases (**Online Methods**). In addition, we examined the Pearson correlation of the gene expression for the two pipelines for each PBMC dataset. We found very high correlations for all datasets for both UMIs and genes (**Supplementary Fig. 13**). The genes per cell output agrees more closely than that of the UMIs per cell (**Supplementary Table 4**), perhaps because of the differences in how the method-specific pipelines process the data.

Supplementary Note 8

Advantages of scumi over existing analysis tools

To systematically analyze and compare data generated from different scRNA-seq protocols and avoid introducing potential biases from method-specific computational pipelines, we developed the scumi software package. Although other scRNA-seq processing pipelines[8, 10-18] (and https://github.com/brwnj/umitools) are available, some are specifically designed to process data from one or a few protocols[10, 11, 15-18] and others can process data from different protocols, but are not scalable to large datasets[8, 12] or are harder to use due to dependency and maintenance issues[14]. scumi is flexible to process reads from different protocols, works with different aligners, is scalable to billions of reads in a single run, has functions for detecting and correcting cell-barcodes, collapsing and filtering UMIs, and efficiently sampling reads to compare different methods, and, most importantly, is relatively easy to use. It takes FASTQ files as inputs and generates gene-cell expression count matrices. Except for removing the reads lacking poly(T) sequences, scumi keeps all other reads in BAM files that can be used for visualization or to pinpoint potential problems in the data. A significant challenge in scRNA-seq analysis, particularly when comparing across methods, is to decide which cells to exclude as low-quality. For samples such as PBMCs and cortex, heterogeneity in the RNA content of the different cell types makes a simple threshold of UMIs or genes per cell inappropriate as it introduces biases against cells with less RNA. Our pipeline provides multiple, customizable solutions to the problem of how to set threshold to filter out low-quality cells, including a mixture model or filtering of initially clustered datasets. Although scumi is not optimized for data generated from any one specific platform, we showed that it produced results that had high concordance with those from method-specific pipelines (**Supplementary Table 4**, **Supplementary Fig. 13**). We

foresee that scumi could be useful for analyzing scRNA-seq data from new platforms and for future benchmarking studies of scRNA-seq methods.

1. Ilicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 29 (2016).
2. Ludwig, L.S. et al. Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339 e1322 (2019).
3. Drokhlyansky, E. et al. The enteric nervous system of the human and mouse colon at a single-cell resolution. *bioRxiv*, 746743 (2019).
4. Muraro, M.J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385-394 e383 (2016).
5. Segerstolpe, A. et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab* **24**, 593-607 (2016).
6. Suryawanshi, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Sci Adv* **4**, eaau4788 (2018).
7. Ziegenhain, C. et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643 e634 (2017).
8. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* **14**, 381-387 (2017).
9. Zhang, X. et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol Cell* **73**, 130-142 e135 (2019).
10. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
11. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
12. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499 (2017).
13. Tian, L. et al. scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput Biol* **14**, e1006361 (2018).
14. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7** (2018).
15. Azizi, E. et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308 e1236 (2018).
16. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol* **20**, 65 (2019).
17. Perugorria, M.J. et al. Histone methyltransferase ASH1 orchestrates fibrogenic gene transcription during myofibroblast transdifferentiation. *Hepatology* **56**, 1129-1139 (2012).
18. Candelli, T. et al. Sharq, a versatile preprocessing and QC pipeline for Single Cell RNA-seq. *bioRxiv*, 250811 (2018).

Supplementary Methods

Cell preparation for the mixture experiment

We purchased human HEK293 (ATCC, #CRL-1573) and murine NIH3T3 (ATCC, #CRL-1658) cell lines and cultured them in Eagle's Minimum Essential Medium (ATCC, #30-2003) supplemented with 10% Heat-Inactivated Fetal Bovine Serum (Thermo Fisher Scientific, #16140071) and in Dulbecco's Modified Eagle's Medium (ATCC, #30-2002) supplemented with 10% Iron Fortified Bovine Calf Serum (ATCC, #30-2030), respectively. We tested the cells for mycoplasma using a PCR-based assay (ATCC, #30-1012K) three days prior to each RNA-seq experiment. We maintained the cell lines at 37 °C, 5% CO_2, following the supplier's recommendations.

We grew HEK293 and NIH3T3 cells to 50-60% confluence, detached them with TrypLE Express Enzyme (Thermo Fisher Scientific, #12604013), quenched them with complete growth medium, and spun them down at 250 x g for 5 min. We washed the cells twice with 1X Phosphate Buffered Saline (PBS), re-suspended them in 1X PBS containing 0.01% Bovine Serum Albumin (BSA, Sigma, #A8806), filtered them through a 40 μm cell strainer (Falcon, #21008-949), and counted them manually. We mixed the HEK293 and NIH3T3 cells at a 1:1 ratio, aliquoted cells for each scRNA-seq method, spun them down again and re-suspended them in a method-appropriate buffer (**Supplementary Table 15**). The passage number and time in culture for the cells are shown in **Supplementary Table 16**.

PBMC preparation

All biospecimens were collected with informed consent by a commercial vendor. Use of all de-identified biospecimens for sequencing at the Broad Institute was further approved by the Broad's Office of Research Subject Protection (ORSP), which determined that the research did

not involve human subjects according to U.S. federal regulations (45CFR46.102f) – determination ORSP-3635. This study complied with all relevant ethical regulations.

We acquired commercially-available PBMCs (AllCells; Lot #PB006F-A5865) as frozen aliquots with 25 million cells for the PBMC1 experiment or isolated them from fresh blood (AllCells) within 4 h of collection using Ficoll-Paque density gradient centrifugation[1] followed by cryopreservation (25 million cells/aliquot) using CryoStor CS10 (StemCell, #07930) for the PBMC2 experiment. On the day of each experiment, we rapidly thawed a PBMC aliquot and transferred the cells to a 50 ml Falcon tube containing 45 ml of RPMI medium without phenol (Thermo Fisher Scientific, #11835-055) containing 2% human AB serum (Corning, #35-060). We spun down the cells at 300 x g for 10 min at 4 °C and resuspended them in the same medium described above at a concentration of 1 x 10^8 cells/ml prior to removing dead cells using the EasySep™ Dead Cell Removal (Annexin V) Kit (StemCell, #17899). We counted the live cells using trypan blue solution (Thermo Fisher Scientific, #15250061) and adjusted the cell concentration to 1000 cells/µl prior to distribution for downstream processes.

Mouse whole cortex dissection

All animal-related work was performed under the guidelines of the Division of Comparative Medicine, with the protocol (# 0416-050-1) approved by the Committee for Animal Care of the Massachusetts Institute of Technology, and was consistent with the Guide for Care and Use of Laboratory Animals (1996 edition). Each cage contained 2–5 mice regardless of genotype; mice were housed at a constant 23 °C in a 12-h light–dark cycle (lights on at 7:00, lights off at 19:00) with ad libitum food and water. We obtained whole cortex samples by euthanizing 1 month old C57BL/6 male mice by cervical dislocation. We then dissected the whole cortex out of the first

hemisphere by a midsagittal cut and a cut made between the cerebellum and the hemisphere. The hippocampus, olfactory bulb, and all basal ganglia were dissected out, and the cortex was placed in 500 µl RNAlater (Thermo Fisher Scientific) in a 1.5 ml tube, followed by incubation overnight at 4 ℃ before storage at -80 ℃. Then second half of the cortex was processed similarly and placed in a different tube. We took only 2 min to harvest the first half of the cortex after the cervical dislocation step and another minute for the second half of the cortex.

Cell sorting

We passed each cell suspension through a 35 µm filter (Falcon) and added 20 µl of 1 µM TO-PRO®-3 Iodide (Thermo Fisher Scientific) per ml to stain dead cells. We used a MoFlo Astrios EQ cell sorter (Beckman Coulter) and set fluorescence activated cell sorting (FACS) gating on forward scatter plot, side scatter plot and on fluorescent channels to pick TO-PRO®-3 negative (live cells) while excluding debris and doublets. We used a 100 µm nozzle to sort single cells into 96-well plates containing 5 µl TCL buffer (Qiagen) with 1% beta-mercaptoethanol for Smart-seq2 and 384-well plates containing 0.6 µl 1% NP40 (Thermo Fisher Scientific) for CEL-Seq2.

Nuclei isolation and sorting

We isolated single nuclei as previously described[2] with the following modifications. We processed the left and right cortical hemispheres of one mouse separately through Dounce homogenization. We then combined the nuclei suspensions during the washing step. We aliquoted ⅓ volume in Tube A and ⅔ volume in Tube B. We spun down the aliquots and re-suspended the pellets in 1 ml DroNc-seq resuspension buffer (1X PBS, 0.01% BSA, 0.16 U/µl

RNase Inhibitor (Clontech/TaKaRa, #2313A)) for Tube A and 1 ml 10X Nuclei resuspension buffer (1X PBS, 1% BSA, 0.2 U/µl RNase Inhibitor) for Tube B. We passed the nuclei suspensions through 20 µm strainers (Celltrics, Sysmex). We counted the nuclei and made the final aliquots for DroNc-seq and sci-RNA-seq from Tubes A and B at a concentration of 300,000 nuclei/ml with the DroNc-seq and 10x resuspension buffers (RBs), respectively. We preserved the rest of Tube A for later bulk RNA extraction and stained the rest of Tube B with Vybrant DyeCycle Violet Stain (Thermo Fisher Scientific) diluted to a final concentration of 10 µM. We used FACS settings similar to those for single cells described above except for picking Violet positive (for nuclei) to sort 10,000 nuclei into 20 µl 10x RB buffer for 10x Chromium (10x Genomics). We also sorted single nuclei into 96-well plates as described above for Smart-seq2.

RNA-seq methods

Smart-seq2

We prepared RNA-seq libraries from single cells and single nuclei sorted into 96-well plates using a modified Smart-seq2 method[3]. Briefly, we cleaned up single cell RNA by adding 11 µl of RNAClean SPRI beads (Beckman Coulter Genomics) into each well of 5 µl of cell or nucleus lysate. After binding and washing twice with 80% ethanol, we eluted the RNA by adding 4 µl of master mix containing 1 µl of 10 mM dNTPs (New England Biolabs), 0.1 µl of 100 µM 3' SMART reverse transcriptase (RT) oligo as 5'-AAGCAGTGGTATCAACGCAGAGTACT(30)VN, 0.1 µl of 40 U/µl RNase Inhibitor (Clontech/TaKaRa), 2.7 µl of 1 M Trehalose (Sigma), and 0.1 µl of 1/2500000x diluted ERCC ExFold RNA Spike-In Mixes (Thermo Fisher Scientific, #4456739). We incubated this mixture

at 72 ℃ for 3 min for denaturing followed by a quick chill on ice and adding the remaining 6 µl reverse transcription mix that contained 2 µl of 5x First Strand Buffer (Thermo Fisher Scientific), 0.1 µl of 100 µM SMARTer II oligo (5'-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G where "+" indicates a locked nucleic acid (LNA) base), 0.25 µl of 40 U/µl RNase Inhibitor, 0.1 µl of 200 U/µl Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific), 3.45 µl of 1 M Trehalose, 0.1 µl of 1 M $MgCl_2$. We incubated the RT reaction at 50 ℃ for 90 min followed by 85 ℃ for 5 min. We then added 15 µl PCR master mix containing 12.5 µl KAPA HiFi HotStart ReadyMix (KAPA Biosystems)) and 0.5 µl 10 µM PCR primer (5'-AAGCAGTGGTATCAACGCAGAGT). We amplified for 21 PCR cycles with conditions as described[3]. We purified PCR products twice with 0.7x AMPure XP SPRI beads (Beckman Coulter Genomics) and eluted in 20 µl EB buffer (Qiagen). We quantified the cDNAs with the Quant-iT™ PicoGreen® dsDNA Assay Kit (Thermo Fisher Scientific) using an EnVision 2104 Multi-label Reader (Perkin Elmer). We normalized the cDNA concentrations and used 0.075 ng of each in a quarter volume of a NexteraXT (Illumina) library construction reaction. We pooled equal volumes from each well and purified with 0.6x AMPure XP SPRI beads.

CEL-Seq2

We prepared CEL-Seq2 libraries from single cells sorted into 384-well plates following the published CEL-Seq2 method[4] with these modifications. 1) To each RT reaction, we added 0.6 µl primer and dNTP mix that contained 4.28 ng RT primer and 1 nmol dNTPs. 2) We used 0.6 µl of NEBNext Second Strand Synthesis Enzyme Mix with 1x NEBNext Second Strand Synthesis Reaction Buffer (New England Biolabs) in a total volume of 12 µl for the 2nd strand cDNA

synthesis. 3) We pooled a total of 1152 µl (12 µl from each well) double stranded cDNA for each 96-well quadrant and used 0.8x volume of SPRI beads mix that contained 76.6 µl of the original RNAClean XP SPRI beads (Beckman Coulter Genomics) and 383 µl 2.5M Sodium Chloride, 20% PEG Solution (Teknova). 4) We used 2.75 µl 10X RNA Fragmentation Buffer (New England Biolabs) in a total 27.5 µl fragmentation reaction volume and later on added 2.75 µl 10x RNA Fragmentation Stop Solution. 5) We performed 15 cycles of PCR and used 0.8x volumes of AMPure XP SPRI beads for the first round of cleanup.

10x Chromium

We loaded 7000 single cells or nuclei onto a Chromium Single Cell 3' Chip A (10x Genomics, PN-120236) and processed them through the Chromium Controller to generate GEMs (Gel Beads in Emulsion). We prepared RNA-seq libraries with the Chromium Single Cell 3' Library & Gel Bead Kit v2 (10x Genomics, PN-120237) following the manufacturer's protocol.
In an additional experiment, we loaded one channel of PBMC1 cells side by side on Chromium™ Single Cell 3' Chip A and Chip B (10x Genomics, PN-1000073) respectively aiming to capture 4000 cells each with Gel Bead Kit v2 and v3 (10x Genomics, PN-1000075) for comparison. We prepared the libraries following the manufacturer's protocols.

Drop-seq

We performed Drop-seq experiments with single cells according to the published paper[5] with the following details. 1) We used a Drop-seq device that generated droplets 125 µm in diameter. 2) We used 100 cells/µl and 120 beads/µl. 3) We flowed cell and bead suspensions at 4 ml/hour, while flowing the oil at 16 ml/hour. 4) We collected the emulsion into a 50 ml Falcon tube for 15

min for each collection and did 2 collections for each experiment. We incubated the collections for up to 45 min at room temperature before breaking the emulsion. 5) We counted the total number of beads collected and made enough PCR reaction mix to have 5000 beads per 50 µl PCR mix in each well. 6) We performed 10 cycles of PCR during the second round of amplification as 98 ℃ for 20 sec, 67 ℃ for 20 sec, and 72 ℃ for 3 min.

DroNc-seq

We performed the DroNc-seq experiment with single nuclei according to the published paper[2] as described for Drop-seq (see above) with the following additional details. 1) We used a DroNc-seq device that generated droplets 75 µm in diameter. 2) We size selected beads <40 µm in diameter using a strainer (PluriSelect, #43-50040-03). 3) We used 300 nuclei/µl and 350 beads/µl. 4) We flowed cell and bead suspensions at 1.5 ml/hour, while flowing the oil at 16 ml/hour. 5) We collected the emulsion into a 50 ml Falcon tube for 22 min for each collection and did 2 collections for each experiment. We incubated the collections for up to 45 min at room temperature before breaking the emulsion.

Seq-Well

We prepared the Seq-Well libraries following the published paper[6] with the following details. 1) We loaded 12,000 cells onto each array to capture 2000 to 3000 cells. 2) We sealed the array with a hydroxylated polycarbonate membrane with pore sizes of 0.01 µm to allow buffer exchange while retaining biological molecules. 3) We used 1500 - 2000 beads per 50 µl PCR reaction volume and pooled 8 PCR products for each final library construction. 4) We purified PCR products with 0.6x AMPure XP SPRI beads followed by another 0.8x SPRI beads clean-up.

5) We prepared Illumina sequencing libraries with 800 pg of pooled cDNA PCR product from 12,000 - 16,000 single-cell transcriptomes attached to microparticles (STAMPs) and purified with 0.6x AMPure XP SPRI beads followed by another 0.8x SPRI beads clean-up.

sci-RNA-seq

For the mixed cell experiments, we started with 1 million cells following the published paper[7] with these modifications and details. 1) We added RNase Inhibitor (Clontech/TaKaRa, #2313B) to all wash buffers in place of DEPC and SUPERaseIn RNase Inhibitor (Ambion). 2) We preloaded a 384-well plate with 1.25 µl of the mix containing 20 µM oligo-dT primer and 2.5 mM dNTP in each well. 3) We counted the fixed cells, re-suspended them at 500 cells/µl, and aliquoted 1000 cells into each well of a 384-well plate. 4) We used RNase Inhibitor in place of RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen) in the first-strand RT reaction. 5) We pooled the cells after RT, then spun them down and re-suspended them in 1 ml PBS buffer containing 1% RNase Inhibitor and 1% BSA prior to staining with NucBlue Fixed Cell ReadyProbes Reagent (Invitrogen) at 2 drops/ml. 6) We sorted 40 cells into each well containing 5 µl TE buffer containing 10mM Tris pH 8.0 and 1 mM EDTA in 96-well plates and gated based on forward scatter plot, side scatter plot, and DAPI signal to avoid doublets. 7) For PCR, we used 1 µl each of the 10 µM P5 and P7 primers and amplified for 22 PCR cycles. 8) We did a two-sided size selection to remove the large genomic DNA (gDNA) fragments by adding 0.5x volume AMPure XP SPRI beads to the PCR products and recovered the supernatant after incubation followed by purifying the supernatant with another 0.5x SPRI clean up.

For single nucleus experiments, we followed the published paper[7] with similar details as for the mixed cell experiment with these additional changes. 1) We eliminated the methanol fixation

step. 2) We used a final concentration of 300 nuclei/µl and aliquoted 600 nuclei/well across one 384-well plate. 3) We pooled the nuclei after RT and filtered through a 35 µm strainer (Falcon) prior to staining and sorting. 4) We gel-purified the final libraries after PCR to efficiently remove the large quantity of gDNA fragments following a procedure similar to a previously published paper[8]. Briefly, we ran the samples on a 10% Criterion TBE gel (Bio-Rad) at 100 V for 65-70 min prior to staining with 1/1000x SYBR Green (Thermo Fisher Scientific) for 10-20 min. We cut out the region from 300 to 600 bp from the gel, collected it into a 1.5 ml Eppendorf tube, crushed it with a disposable pestle (Kontes/Kimble Chase, #749520-0090), added 400 µl of 0.3M NaCl to the crushed gel pieces and incubated with rotation at room temperature for at least 4 h to overnight. We used a Spin-X cellulose acetate filter (Corning, #8163) to remove the gel particles by spinning for 3 min at 15,000 x g. We ethanol precipitated the flow through by adding 1 µl glycogen (20 µg/µl; Roche, #10901393001) and 2.5 volumes 100% ice cold ethanol followed by incubation at -20 °C for at least 30 min and then 20 min of centrifugation at 15,000 x g and 4 °C. We washed the pellets with 1 ml 70% ethanol and air dried before resuspending the libraries in EB buffer (Qiagen).

inDrops

We prepared inDrops libraries based on a previously described protocol[9, 10] at the Single Cell Core at Harvard Medical School with 2000 to 2500 cells collected for each library. We made the following modifications in the primer sequences to eliminate the need for custom sequencing primers. 1) We changed the sequences of RT primers on hydrogel beads to: 5'-CGATTGATCAACGTAATACGACTCACTATAGGGTGTCGGGTGCAG[barcode1,8nt]GTC TCGTGGGCTCGGAGATGTGTATAAGAGACAG[barcode2,8nt]NNNNNNTTTTTTTTTTT

TTTTTTTV- 3'. 2) We changed the PE2-N6 primer sequence in step 151 of the published protocol[10] to: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNN-3'. 3) We moved the indexing barcode from PE1 primer to PE2 in steps 157 and 160 of the protocol[10] as the following: PE1-5'- CAAGCAGAAGACGGCATACGAGATGGGTGTCGGGTGCAG-3', PE2-5'-AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC-3', where XXXXXX is the indexing barcodes for multiplexing libraries.

RNA-seq from bulk samples

We pelleted the cultured cells, PBMCs, and isolated nuclei leftover from Tube A after taking out the aliquot for DroNc-seq (see Nuclei isolation and sorting section) by spinning at 500 x g for 5 min at 4 ℃. After removing the supernatant, we resuspended them in 100 µl DNA/RNA Shield (Zymo Research) and stored them at -80 ℃. We extracted total RNA using Quick-RNA Mini Prep (Zymo Research) and followed the vendor protocol with a DNase treatment step. We prepared the RNA-seq libraries using the TruSeq® Stranded mRNA Library Prep Kit (Illumina) following the manufacturer's protocol except for the following modifications. 1) We did the elution, priming and fragmentation of mRNA at 85 ℃ for 4 min. 2) We used SuperScript III (Thermo Fisher Scientific) in place of SuperScript II and performed reverse transcription at 55 ℃. 3) We used a different set of barcoding indices rather than those in the TruSeq kit for the ligation and final PCR steps.

Sequencing

To avoid batch effects due to varying sequencing quality among Illumina flowcell lanes, for each comparison experiment, we pooled together the single cell or nucleus libraries from different

methods with normalized molar concentrations and loaded them onto a HiSeq2500 (Illumina) in the 8-lane high output or 2-lane rapid run mode. The inDrops libraries were sequenced separately because they have an opposite read structure from those generated from the other methods with the cDNA in read 1 needing a long read and indexing information in read 2 needing a short read. We performed additional sequencing for some libraries in an attempt to sequence similar numbers of reads per cell for each low or high throughput method. We aimed for 50,000 to 100,000 reads per cell for high-throughput methods and 750,000 to 1,000,000 reads per cell for low-throughput methods.

For analysis, we trimmed the barcode reads (including inline barcodes from Read1 or Read2 and also the index read) back to the original length required for each library type, since the same run parameters had to be used throughout the sequencing run. Additional details are in **Supplementary Tables 9** and **17**. We sequenced the bulk RNA-seq libraries on a NextSeq500 (Illumina) with 75 bases each for reads 1 and 2 and 8 bases each for index reads 1 and 2.

Read demultiplexing

As methods using a single index read (e.g., 10x Chromium) or dual index reads (e.g., Smart-seq2) were pooled together for sequencing, we adopted a two-round demultiplexing strategy. We first demultiplexed the reads using both index reads to extract reads from Smart-seq2 and sci-RNA-seq. In the next round, we further demultiplexed the reads using a single index read. In the case of barcode collisions (i.e., a single index library index read matches that of a dual index library), we extracted the single index library reads from the unmatched reads after the first demultiplexing step. We only kept reads with one or no mismatches in the index reads.

Multiplet Detection

In the mixture experiments, we considered cells multiplets if they had >15% of UMIs from mouse and >15% of UMIs from human. For Smart-seq2, this calculation was instead based on the percent reads, though with the same cutoffs. We report the observed multiplet rate, though the actual multiplet rate was likely to be about twice this frequency due to the presence of undetected multiplets that are all from cells of the same species (assuming about equal numbers of human and mouse cells).

Unmapped reads

We extracted unmapped reads from the corresponding BAM files for each method from the PBMC datasets. We used Trimmomatic[11] v0.39 to trim and filter these reads, based on QC metrics (see below), polyA content, known adapters, or all three. In all cases, we used MINLEN:25. For those with QC, we used LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15, while for those filtered based on poly(A) or adapter content we used ILLUMINACLIP:${adapter}:2:30:10, where adapter was the FASTA file containing the associated primers and polyA sequences (**Supplementary Table 18**). Reads trimmed with all three filters were then re-mapped using the same methods and references as in the original PBMC analysis.

Antisense reads

To ensure we could determine which transcript antisense reads mapped to, we removed all transcripts in the mm10 GTF that overlapped any other transcript either on the same or opposite strand. We then extracted all reads from the scumi-generated BAM files that mapped antisense to

one of these transcripts (either in the introns or exons) using bedtools "intersect"[12], and picked 1,000,000 of these reads at random for each method and each dataset (except picking 10,000 reads per well for sci-RNA-seq) to reduce computational time. We excluded Smart-seq2 as it is not a strand-specific method. From these reads, we generated a bed file with one entry per read, corresponding to the 500 bp adjacent to the 3' end of the read in the genome, and extracted the corresponding sequences from the mm10 FASTA file using bedtools "getfasta"[12]. We then iterated over each position in each of these sequences, and recorded which 5-mer occurred at that position in that read, thus counting the number of times each 5-mer occurred in each position relative to the 3' end of each antisense read. We then extracted the information about the AAAAA kmer for our analysis.

Detection of both human and mouse genes expressed in single cells

In the Mixture experiments, for each human cell, we computed the percentage of detected mouse genes in that cell. Similarly, we repeated this analysis to detect human genes from mouse cells. To show the relationship between sequencing depth and the number of detected genes from the other species, we drew scatter plots for the detected human cells and mouse cells where the *x*-axis is the number of detected human genes and the *y*-axis is the number of detected mouse genes. Generally, the number of detected genes *y* from the other species increases linearly with the number of detected genes *x* from the called species. Therefore, we used robust linear regression (Huber M-estimator, implemented in the MASS R package) to fit $y = ax + b$, for human cells and mouse cells, separately. The fitted line and its slope (a) were plotted on the scatter plots.

Pseudo-bulk expression profile comparisons

We built pseudo-bulks by combining the expression vectors from all the human cells (or mouse cells) in the mixture experiments and computed the Pearson correlation coefficients between the pseudo-bulks from different methods. The pseudo-bulk gene expression profile for a set of cells was computed by first taking the sum of the gene expression vectors from these cells, where the gene expression vector of a cell was the raw UMI count vector, one element for a gene. The dimensionality of a gene expression vector was the number of genes. To make the bulk gene expression vectors from different sets of cells comparable, a bulk gene expression vector was normalized by dividing the total number of UMIs from all the cells used in computing that bulk gene expression vector, and further multiplying by 10^4 and finally taking the natural log transform (adding one before the log transformation to make all the elements of the bulk vector positive). For Smart-seq2 data, we summed the read counts from cells and then transformed the counts to transcripts per 10K (TP10K) values ($10^4 (N_i / L_i) / \text{sum}_i (N_i / L_i)$, where N_i is the total number of counts for gene *i* and L_i is the length of gene *i*) to get the pseudo-bulk expression.

Bulk RNA-seq processing and analysis

Expression in bulk RNA-seq data was quantified with RSEM[13]. We extracted TPM values and divided by 100 to transform the data into TP10K. For PBMCs, the log of the resulting TP10K was then compared to the pseudo-bulk data from each of the single cell methods (see Pseudo-bulk expression profile comparisons section) using Pearson correlation, with the exception of Smart-seq2. For Smart-seq2, we instead combined all FASTQ files into one joint FASTQ, and quantified TPM using RSEM using the same pipeline as for the bulk data. The percent

mitochondrial transcripts for bulk data was calculated by taking the sum of TPM values over all mitochondrial genes (defined above) and dividing through by 1,000,000.

In addition to expression, *de novo* transcript construction was performed on the bulk RNA-seq data in the hope of correcting mistakes in the standard annotation. To achieve this reannotation, bulk RNA-seq data was mapped to the genome with HISAT2[14] v 2.0.5, using the –dta flag and otherwise default parameters. *De novo* transcript construction was then performed with StringTie[15] v0.1.18 with the parameters -v --rf -c 10 and –G, using the GTF from the scumi pipeline.

In order to quantify the improvement in performance by using the new annotation, we extracted reads from the annotated BAM file output by scumi that did not overlap any exons in the standard annotation (using the equally sampled data). We used featureCounts[16] to count the number of these reads that overlapped an exon in the reannotated GTF output, to quantify how many more reads per cell we recovered by using the reannotated reference.

To analyze the variation in single-cell profiles, we computed the Spearman correlation of each cell profile in each of the Mixture libraries with the corresponding bulk sample gene expression profile. For the bulk data, we used the TPM values; for single-cell profiles, we used the UMI values as the Spearman correlation coefficient is invariant with a monotonic increasing transformation of the input data.

Technical precision

We computed the extra Poisson variations that include both biological variations (assumed to be small and the same across methods in our data) and technical variations that cannot be explained by simple Poisson models[17]. High extra Poisson variations indicate either high technical

variation or mispecification of the Poisson model (as seen with Smart-seq2). Specifically, let the vector $\mathbf{x}=[x_1,\ldots,x_N]^T$ denote the expression of a gene x across N cells, where x_i is the expression of that gene in cell i. We assume that x_i follows a Poisson distribution with parameter λe_i, where λ corresponds to the number of transcripts present in this specific gene and e_i accounts for the sequencing depth variation of cell i. For simplicity, we assume that e_i is the same for all cells after median normalizing the UMI counts, i.e., $x'_i=x_i/s_i*\text{median}(\mathbf{s})$ follows a Poisson distribution with parameter λ'. Here $\mathbf{s}=[s_1,\ldots,s_M]^T$, where s_i denotes the total number of UMIs across genes in cell i, and M is the total number of cells from a method. We then computed the extra Poisson variations defined as the empirical coefficient of variation (CV) subtracted from the Poisson CV. High extra Poisson CV indicates high technical noise. We used only genes expressed in more than 5% of the cells.

Sensitivity analysis of published datasets

For each published dataset, we downloaded the FASTQ files, and processed them through scumi with the following small modifications. 1) We did not use the poly(T) filter as those datasets did not sequence that part of the libraries. 2) For Smart-seq2, the reads were single end (not paired-end as in our study), so that we removed the paired-end flags from the pipeline. 3) For CEL-Seq2, the downloaded data had a slightly different UMI and cell barcode structure, so that we modified the configuration file accordingly. 4) We required scumi to return 10,000 cells per library for Drop-seq and 10x Chromium datasets. For CEL-Seq2 and Smart-seq2 we used the same cell filtering approach as for the mixture data, while for the other datasets we used the approach as for the PBMC data. We sampled the reads so that each pair of samples had roughly

the same number of reads per cell, re-ran the scumi pipeline and cell filtering steps on the sampled data, and compared the number of genes per cell between methods.

Effects of sequencing depth on gene detection

To estimate the sensitivity of each method at different sequencing depths, we sampled datasets to a given numbers of reads per cell. To efficiently sample without running scumi starting over using FASTQ files as inputs, we instead sampled the 'molecular_info.h5' matrix output from running scumi. The molecular_info.h5 matrix recorded the number of reads for each UMI of a gene in each selected cell. Given N, the total number of reads in the input FASTQ files, C, the number of detected cells, M, the total number of reads recorded in the 'molecular_info.h5' matrix, and R, the expected number of reads per cell after sampling, we needed to draw $C*M*R/N$ reads recorded in the 'molecular_info.h5' matrix. This sampled molecular_info.h5 matrix was converted to the final UMI matrix after collapsing UMIs. The code for sampling is part of the scumi package.

To compare results from sampling the raw reads rather than molecular matrices, we used five of the eight datasets from PBMC1. We sampled the raw reads such that the number of reads per cell was roughly the same across the data from different methods. We further sampled the molecular matrices from the original, non-sampled data such that the number of reads per cell after sampling matched that from sampling the raw reads. We then compared the number of genes detected from each cell obtained from the sampling strategies.

Detection of differentially expressed genes between methods and between replicates

For the Mixture datasets for each method, we separated mouse and human cells and filtered out reads aligning to mouse genes from the human cells and vice versa.

For differential expression analyses between methods, for each method, we compared gene expression between cells from that method and all other methods using logistic regression (where the dependent variable was 0 if a gene was not expressed, 1 if it was), including covariates for the number of genes and for experimental batch (Mixture1 vs. Mixture2). We preformed this analysis for human and mouse cells separately and only considered genes that were expressed in at least 5% of cells. We used UMI counts for all the methods, except for Smart-seq2, where we used read counts.

For each method, we picked the 100 most upregulated genes (smallest p-values and odds ratio > log(2)), and calculated the average GC content and length for all transcripts of each gene.

For differential expression analyses for the same method between replicates, we followed a similar strategy, as above, separately for human and mouse cells. For each method and each gene, we compared between Mixture1 and Mixture2 using logistic regression, correcting for nGene, the number of genes detected in each cell. We used DropletUtils[18] (v1.0.3) to sample each dataset so that Mixture1 and Mixture2 had the same average number of UMIs per cell before performing our comparison.

For investigating the expression of immediate early genes, we extracted data for the genes identified as being induced by dissociation[19]. We scored each cell by calculating the natural log of the sum of the TP10K value for these immediate early genes.

Parameter selection for clustering analysis

We considered the following parameters for the Louvain clustering: 1) We tested different numbers of principal components (20, 30, 50, 100). 2) We varied the number of neighbors *k* used

in building the *k*-NN graph. This parameter was dependent on the number of cells from each cell type and was either 15 or 30 for high-throughput datasets from PBMCs. The only exception was the Seq-Well dataset from PBMC2 which had only 551 cells and *k* was of 10, 15, or 30. For the data from low-throughput methods Smart-seq2 (from PBMC1, PBMC2, Cortex1, and Cortex2) and CEL-Seq2 (from both PBMC1 and PBMC2), the parameter was either 5, 10, or 15. 3) We varied whether we used only variable or all genes for clustering. 4) We varied the resolution parameter that could influence the number of clusters (high-throughput methods: 0.8, 1.0, 1.2, 1.5; low-throughput methods: 0.5, 0.8, 1.0, 1.2, 1.5). For the cortex datasets, the search space of parameters was the same as for the PBMC2 Seq-Well dataset with the exception of DroNc-seq data from Cortex2, which had only 892 cells so its parameter search space was the same as for Smart-seq2 data. We then scored each clustering by the formula $\sum_m (\overline{\text{AUC}}_m - 0.7)$, where $\overline{\text{AUC}}_m$ is the average AUC for cell type *m* (the average was over the clusters assigned to cell type *m*). Therefore, if cells of a single type, e.g., the B cells were partitioned into several clusters, their AUC average was smaller compared to the case where all B cells were grouped into one cluster. For this reason, the average AUC penalized partitioning cells of the same type across several clusters. The summation over all cell types in the formula favored the recovery of more distinct cell types (with AUC greater than 0.7). The actual parameters used for clustering are listed in **Supplementary Table 19**.

Scoring each dataset to compare scRNA-seq methods

We used the AUCs to quantify the quality of each recovered cell type. After clustering analysis and assigning cell types to clusters, we also calculated the AUC of each cell type (from its cell type marker gene scores in separating that cluster of cells from the rest).

Standard, method-specific computational pipelines

We processed each of the PBMC datasets through the standard pipeline for that scRNA-seq method and compared the output statistics to those from scumi. If there were multiple sequencing runs for a given library, we combined the FASTQ files before further analysis. Additionally, files were renamed to match the naming schemes required for a given pipeline. For 10x Chromium v2 data, we used the Cell Ranger pipeline[20] (v2.0.0) with the same reference genome (stored in a FASTA file) and GTF file that were used for scumi. We truncated index reads to 8 bases if the original sequence was longer. We then ran the Cell Ranger 'count' command with the –nosecondary flag and with –cells set to the expected number of cells. The gene by cell UMI count matrix was then extracted, with no additional cell filtering. For 10x Chromium (v3), we used a similar pipeline, except we used the Cell Ranger (v.3.0.2) pipeline with the –expect-cells flag instead of the –cells flag, and did not truncate the index reads.
For Drop-seq and Seq-Well data, we used the Drop-seq pipeline[5] (v1.3, available at https://github.com/broadinstitute/Drop-seq). We created Drop-seq references from the GTF and FASTA files used in the scumi reference with the make_resource_bundle.py script in the Drop-seq pipeline. We processed the reads with the Drop-seq pipeline with the number of core barcodes set to the expected number of cells and using the reference we generated, though otherwise the same parameters as in Shekhar et al.[3]. To select the number of cells, we used three steps: first, we applied the cell_selection.R script in the Drop-seq pipeline to the output of the

standard Drop-seq pipeline, feeding in the expected number of cells. Second, we used the DigitalExpression command from the Drop-seq pipeline to create a digital expression matrix using the cells selected. Finally, we chose a cutoff by visual inspection of the histogram of genes per cell.

We used the inDrops pipeline from the initial publication[9] (available at https://github.com/indrops/indrops). We first constructed references according to the inDrops README using RSEM[13] (v1.3) with the GTF and FASTA files used to construct our scumi references. We trimmed the reads to be the same length in all runs (read 1 50 bases, read 2 19 bases, index reads 8 bases). We then ran the pipeline using the default YAML configuration, modified to work with Bowtie[21] (v1.1.1). We then extracted the gene by cell UMI count matrix. For cell selection, we chose the nGene cutoff used by visual inspection of the histogram of nGenes.

For CEL-Seq2 data, we used the pipeline from the original publication[4] (available at https://github.com/yanailab/CEL-Seq-pipeline). We created a Bowtie 2 reference using the bowtie2-build command in Bowtie 2[22] (v2.2.1), using the same FASTA file as for the other pipelines. We processed the reads with the CEL-Seq2 pipeline with the default configuration file, modified to include our updated reference. We extracted and combined the gene by cell UMI count matrix. For cell selection, we chose the nGene cutoff used by visual inspection of the histogram of nGenes.

For Smart-seq2, we used the same pipeline as in the scumi results – HISAT2[14] to map the FASTQ file, featureCounts[16] to annotate the BAM file, followed by counting reads that overlap the exons of a unique gene – except for cell selection we chose the nGene cutoff by visual inspection of the histogram of nGenes.

Clustering combined datasets

For the PBMC and cortex datasets, to find cell types that were missed in analyzing each dataset independently, we took the count data for each method (after sampling the same number of reads and cells as describe above) in each experiment, and merged them into one count matrix. For Smart-seq2 we used reads counts, while for the other methods we used UMI counts. We loaded the results into R using Seurat, normalized to log counts per 10,000, and scaled the data. We identified variable genes separately for each library using the 'FindVariableGenes' function (with x.low.cutoff=.1 and x.high.cutoff=5), and keeping genes that were variable in at least 5 libraries for the PBMCs or at least 3 libraries for the Cortex. We then performed PCA with those genes, and used the results for Harmony[23] with parameters, theta = 4 (for PBMCs) or theta=6 (for cortex), nclust = 50, and max.iter.cluster = 100. The resulting 20 Harmony components were used for clustering and t-SNE visualization with default parameters. Cell type annotations were based on the PBMC and cortex marker genes. For cortex, we also manually identified two clusters, one that seemed to correspond to the unidentified cells in sci-RNA-seq, and the other that seemed to be fibroblast-like cells (not considered in our analysis), both of which we labelled as unassigned.

To check the Harmony cell type results, the data for each library were extracted from the Harmony object one by one, and the cell type clusters identified in the combined Harmony analysis were reannotated using the AUC cell type identification method with only cells from that one library (excluding unassigned cells). This updated annotation was compared to the combined Harmony annotation to determine whether the same cell types were identified.

To assess if combining data from different single cell methods led to better clustering, we extracted the cell type labels for each 10x Chromium (v2 or v3) dataset one by one from the Harmony clustering and calculated the associated AUC for each cell type. We then compared the resulting AUCs to the corresponding AUCs from the cell types calculated on each dataset independently.

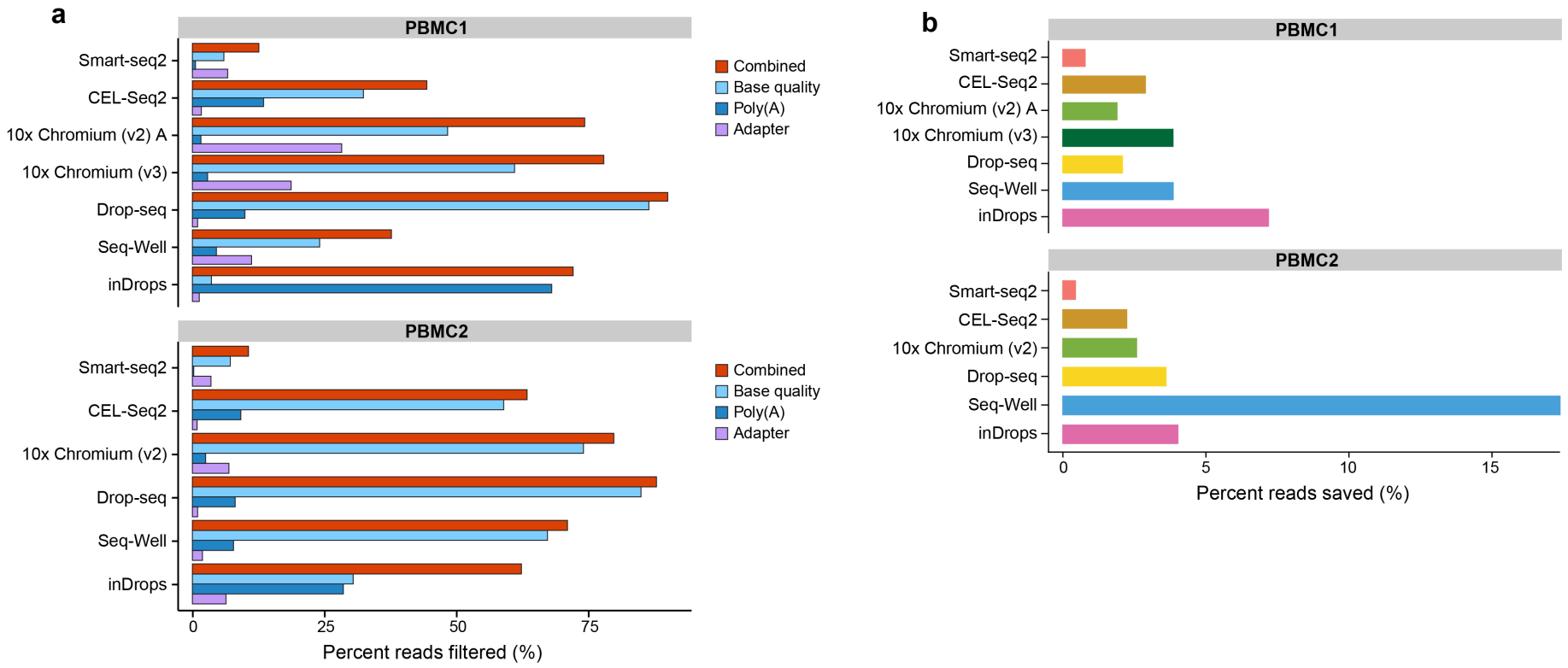
Statistics

For testing for differentially expressed genes between methods and batches, we fit a logistic regression model, including number of gene, batch, and (where appropriate) method as covariates, using the glm function in R. We performed a 2-sided *t*-test on the coefficients of interest, and performed FDR correction with p.adjust. For testing for differences in the Immediate Early gene score, we fit a linear model with the lm function in R, with number of genes and batch as covariates, and we used a 2-sided *t*-test to test the effects of batch on the Immediate Early gene score. These tests treat each cell as an independent observation. The number of human or mouse cells in each sample is reported in **Supplementary Table 3**.

1. Lee, J. et al. Universal process-inert encoding architecture for polymer microparticles. *Nat Mater* **13**, 524-529 (2014).
2. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* **14**, 955-958 (2017).
3. Shekhar, K. et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323 e1330 (2016).
4. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 77 (2016).
5. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
6. Gierahn, T.M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* **14**, 395-398 (2017).
7. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
8. Levin, J.Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709-715 (2010).

9. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
10. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* **12**, 44-73 (2017).
11. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
12. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
13. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
14. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
15. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295 (2015).
16. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
17. Ziegenhain, C. et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643 e634 (2017).
18. Griffiths, J.A., Richard, A.C., Bach, K., Lun, A.T.L. & Marioni, J.C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature communications* **9**, 2667 (2018).
19. van den Brink, S.C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat Methods* **14**, 935-936 (2017).
20. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
21. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
22. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
23. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* (2019).

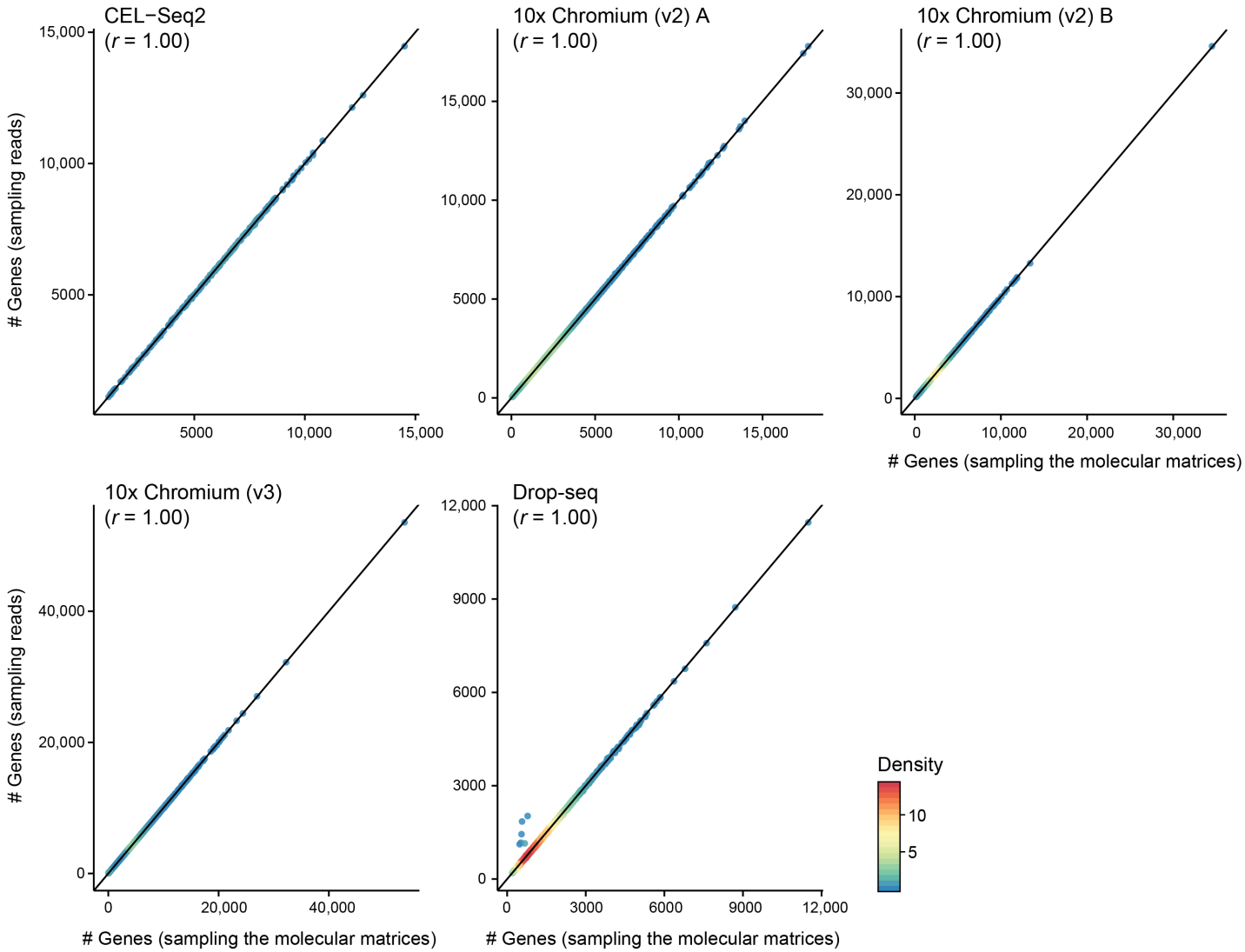
Supplementary Figure 1



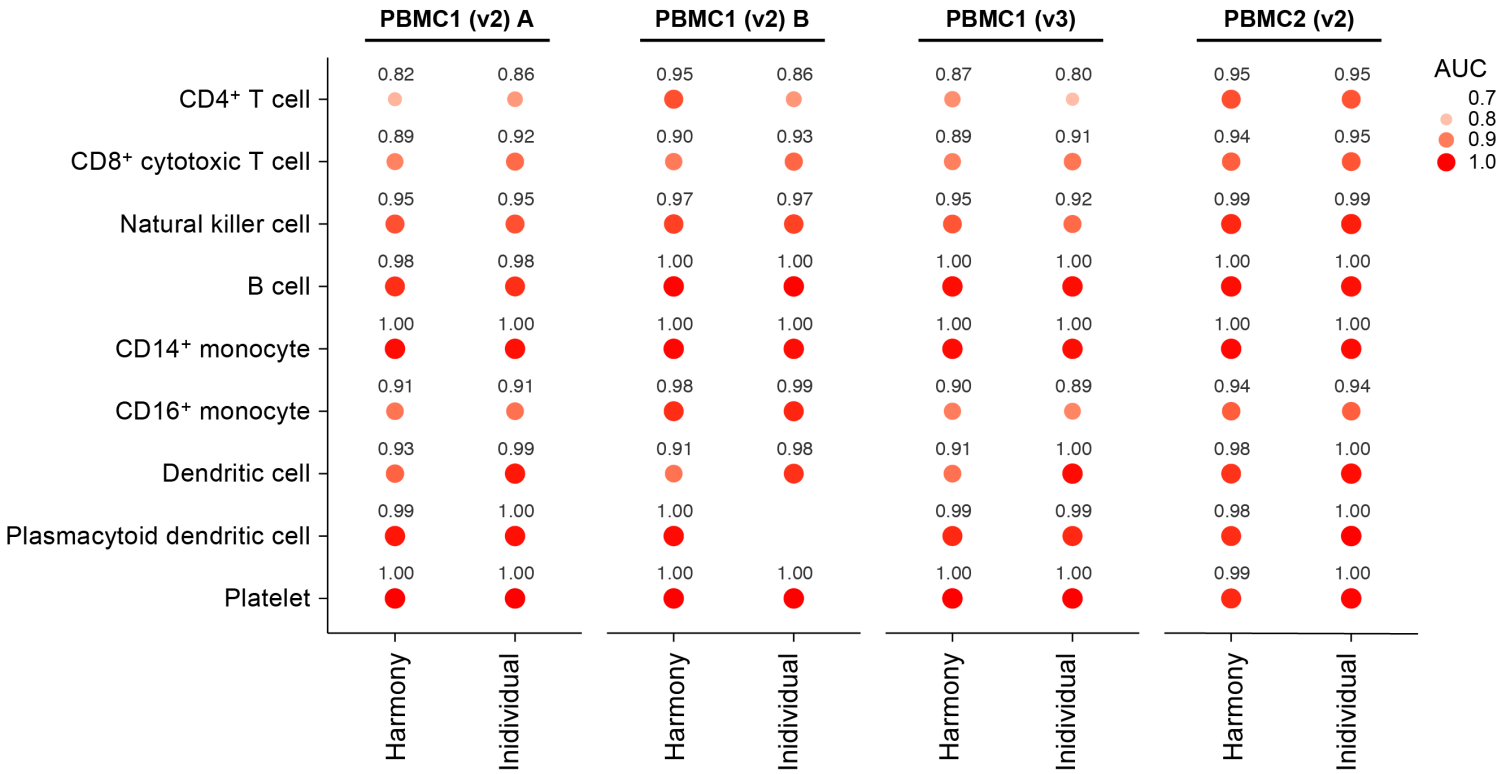
Supplementary Figure 1. **Properties of unmapped reads.**

(**a**) For each method, the percentage of reads filtered as low base quality, adapter, poly(A), or a combination of all three. (**b**) The proportion of previously unmapped reads successfully re-mapped to the genome following trimming for each method.

Supplementary Figure 2

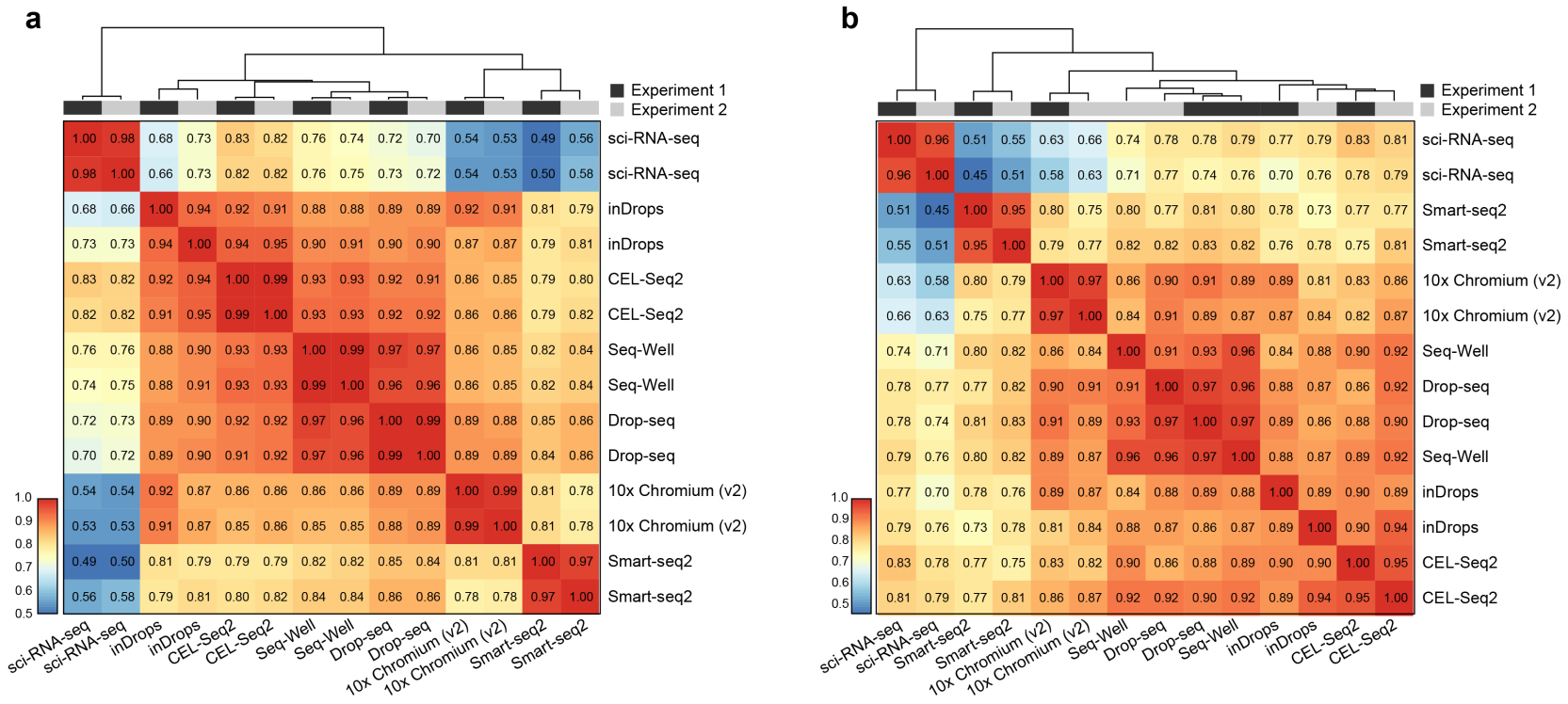


Supplementary Figure 2. **Indistinguishable number of detected genes by sampling raw reads or molecular matrices.**
Number of genes detected when sampling raw reads (*y*-axis) or molecular matrices (*x*-axis) across the five libraries. Pearson correlation coefficients (*r*) are in the upper left corners (*n* = 1 biologically independent sample per scatter plot). Color encodes the density (two-dimensional kernel density) estimate at each point.



Supplementary Figure 3. **Comparison of AUC scores for clustering individually or with other datasets.**

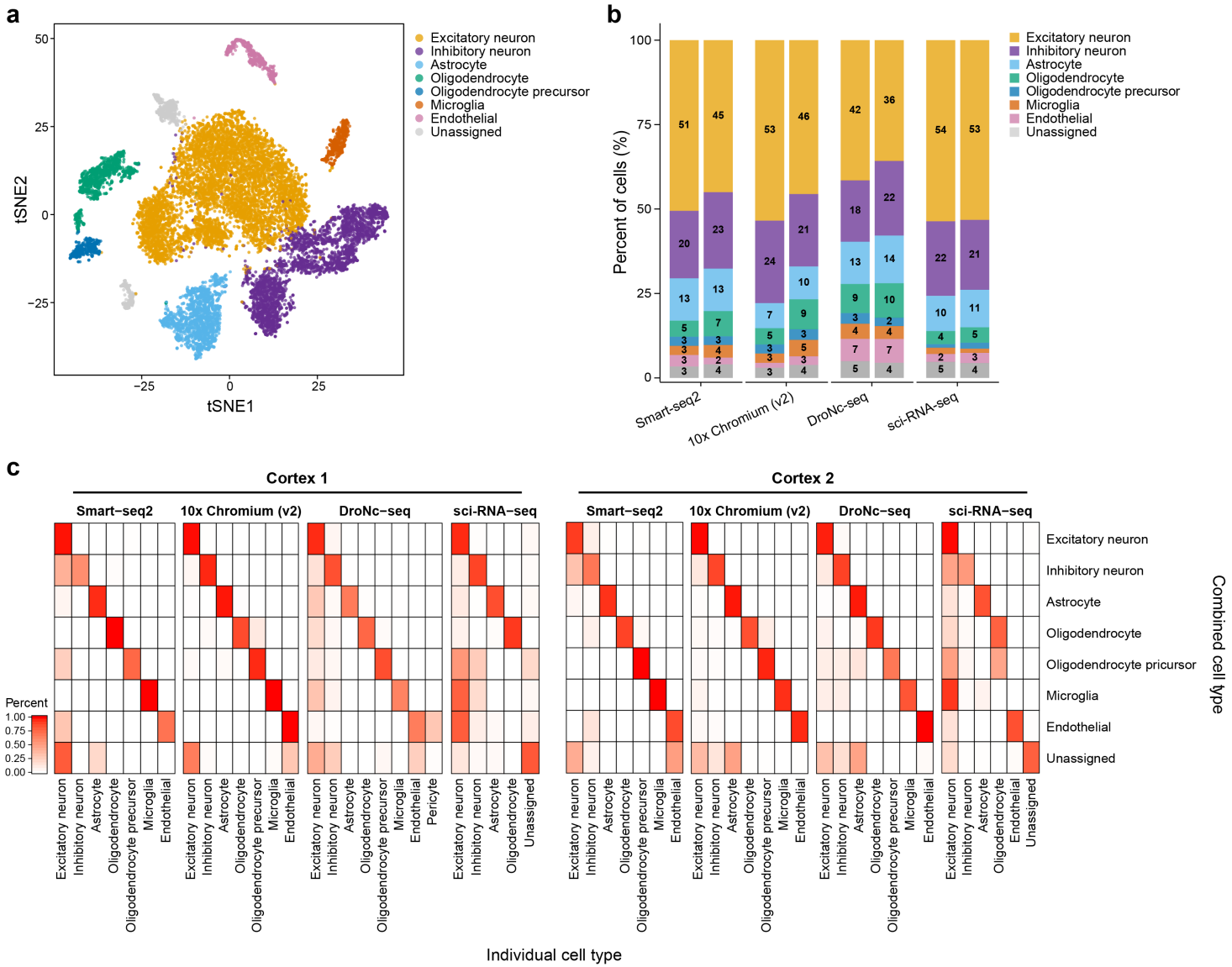
AUC scores for the 10x Chromium PBMC datasets calculated by individually clustering and by clustering with other non-10x Chromium datasets using Harmony (*n* = 2 biologically independent samples).



Supplementary Figure 4. Correlation of pseudo-bulk expression profiles.

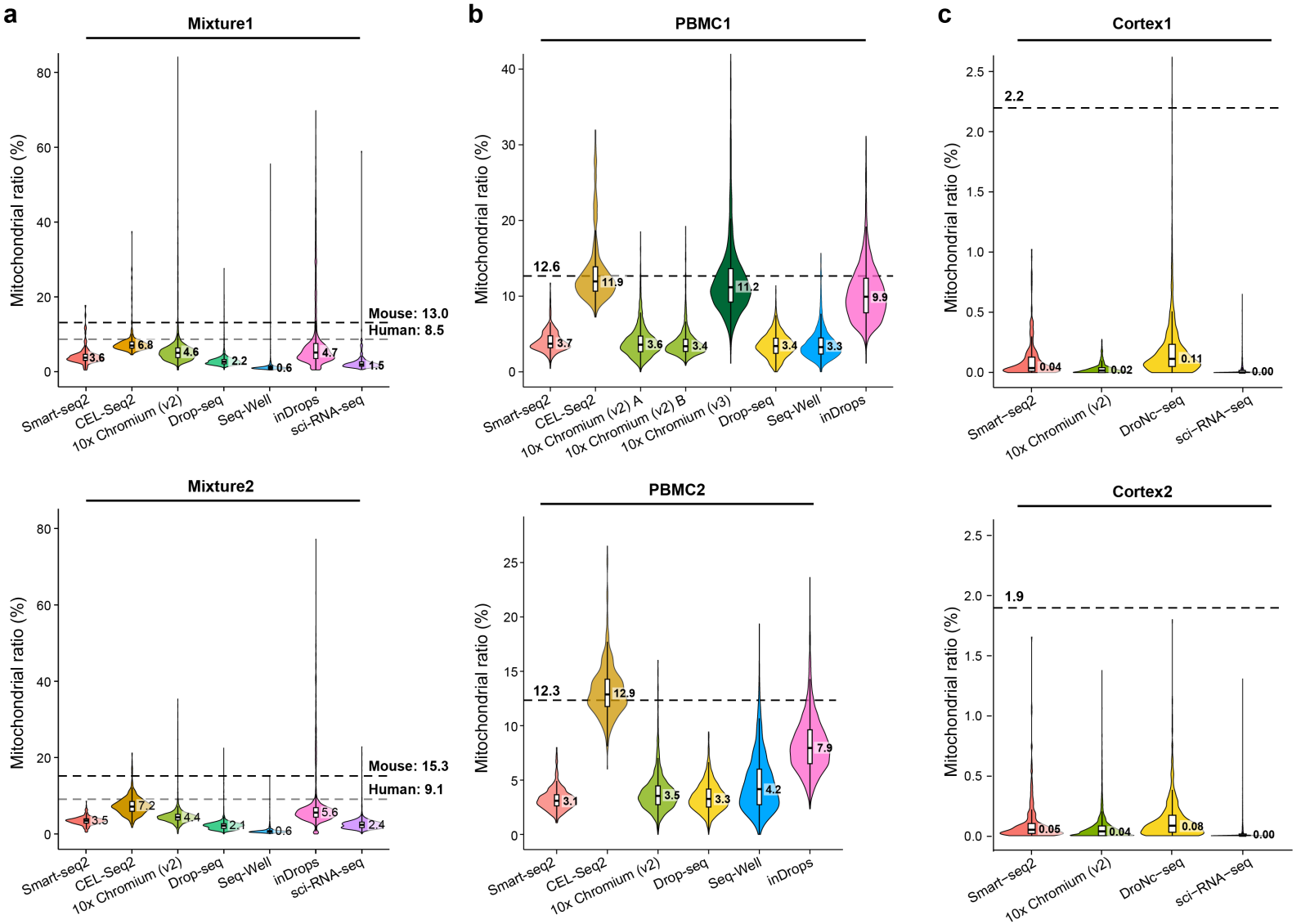
The Pearson correlation coefficients among pseudo-bulk profiles of (**a**) human and (**b**) mouse cells obtained from different methods (*n* = 2 biologically independent samples per panel). High correlations indicate high reproducibility. The number in each cell represents the Pearson correlation coefficient between two pseudo-bulks. The Pearson correlation coefficient matrices were clustered using complete-linkage hierarchical clustering. The expression vectors of cells from a method were added together to form a pseudo-bulk profile. Human cells and mouse cells were analyzed separately.

Supplementary Figure 5



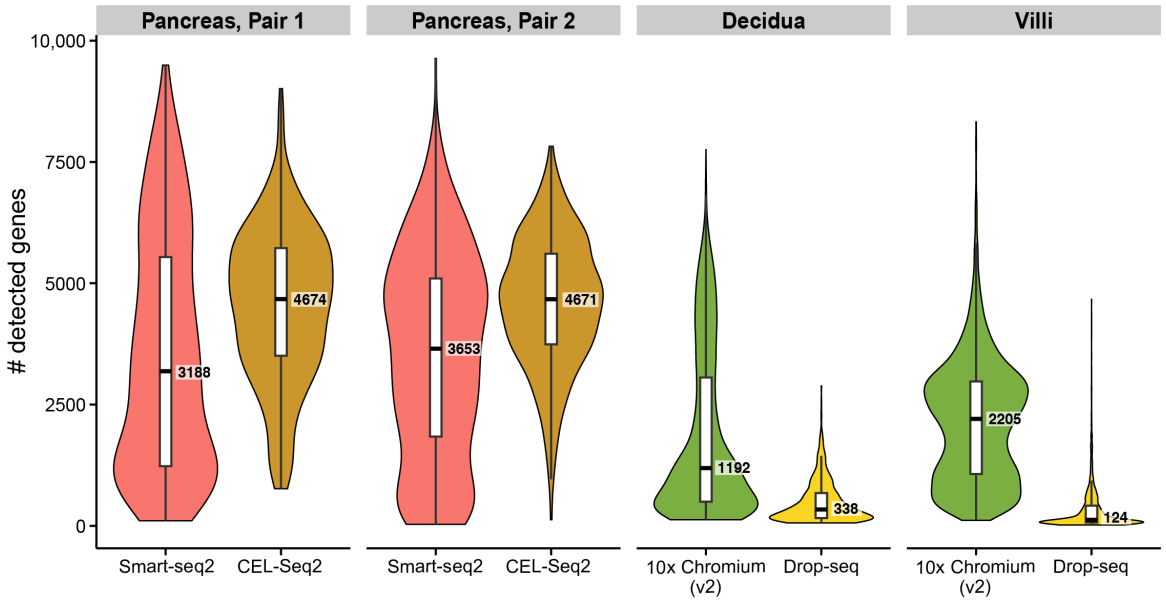
Supplementary Figure 5. **Analysis of combined cortex datasets.**
(**a**) t-SNE plot generated with Harmony clustering all cortex nuclei in this study (*n* = 2 biologically independent samples). (**b**) Cell types identified from each experiment, for each method. All libraries contain nuclei of every cell type, according to this joint annotation. (**c**) For each annotated cell type and method in the jointly clustered dataset (*y*-axis), we calculated the percentage of nuclei from that cell type that come from each cell type in the individual level clustering results (*x*-axis).

Supplementary Figure 6



Supplementary Figure 6. Proportion of mitochondrial reads.

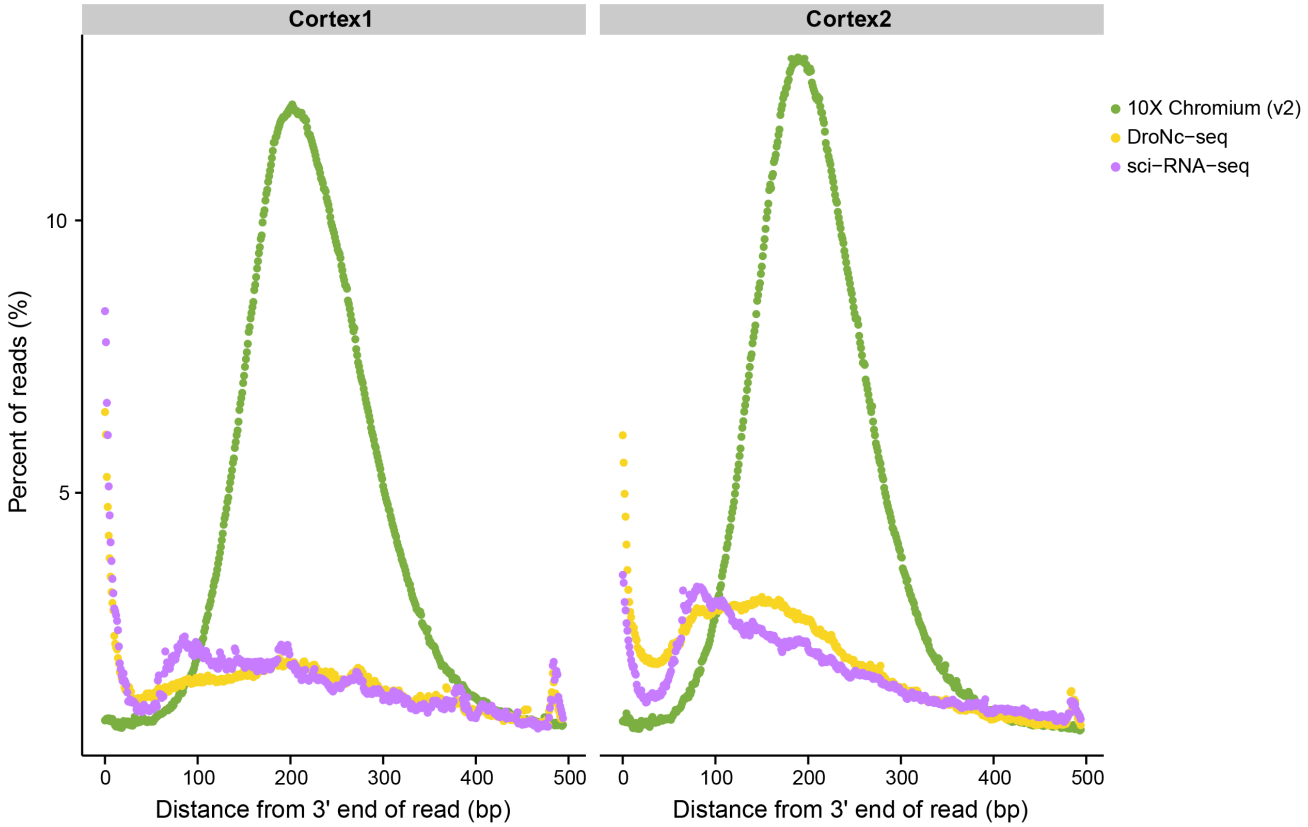
(**a**) Human and mouse cell line mixtures, (**b**) PBMCs, and (**c**) cortex. The mitochondrial ratio distributions in cells from each method (*n* = 1 biologically independent sample per plot). Dotted line and adjacent number: mitochondrial ratio for corresponding bulk dataset(s). Violin and box plot elements are defined as in **Fig. 2**.



Supplementary Figure 7. Sensitivity for public datasets.

Distribution of the number of genes detected per cell (*y*-axis) for scRNA-seq methods from pancreas, placental villi, and placental decidua (*n* = 1 biologically independent sample per tissue region per violin plot). See **Supplementary Table 12** for the numbers of cells used. Violin and box plot elements are defined as in **Fig. 2**.

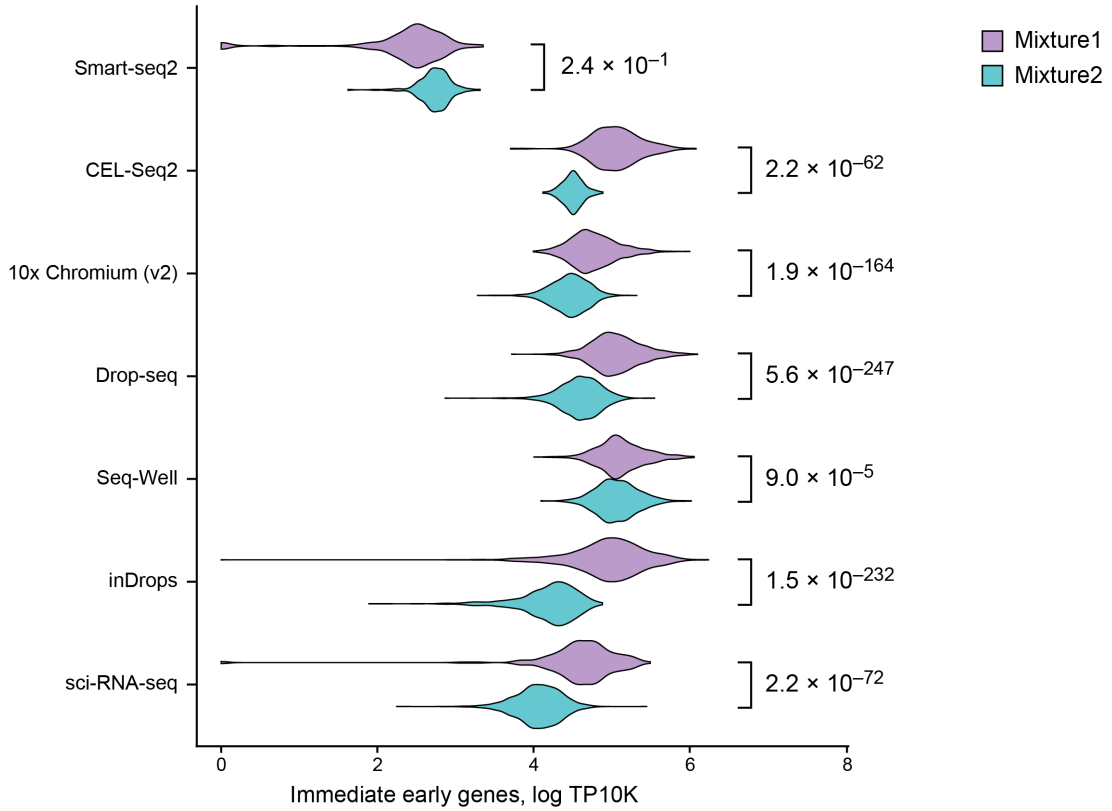
Supplementary Figure 8



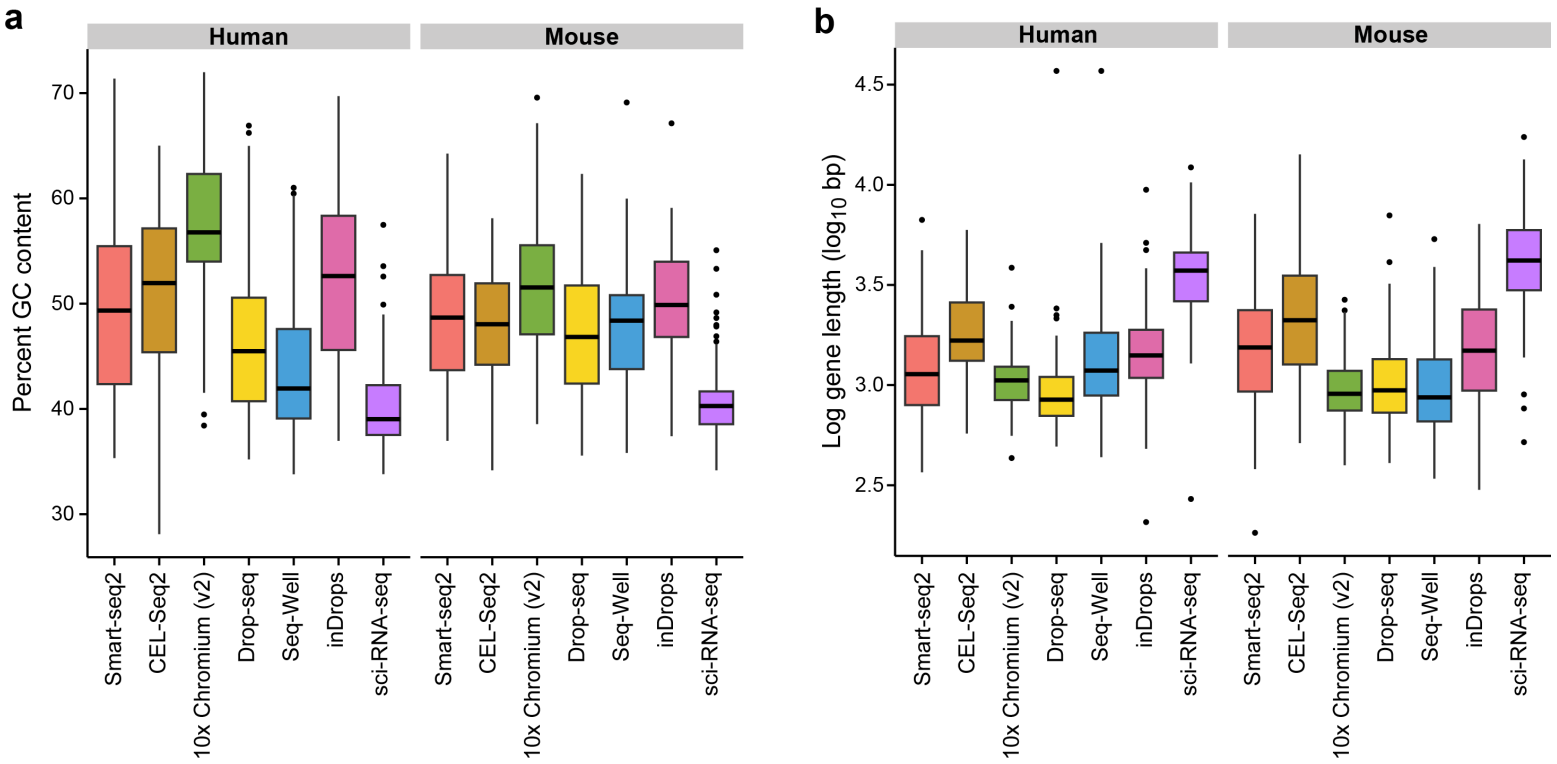
Supplementary Figure 8. Fraction of reads with AAAAA sequence adjacent to the 3' end of antisense reads.

Percent of reads (*y*-axis) with AAAAA sequence at different distances (*x*-axis) from the 3' end of antisense reads in Cortex1 and Cortex2.

Supplementary Figure 9



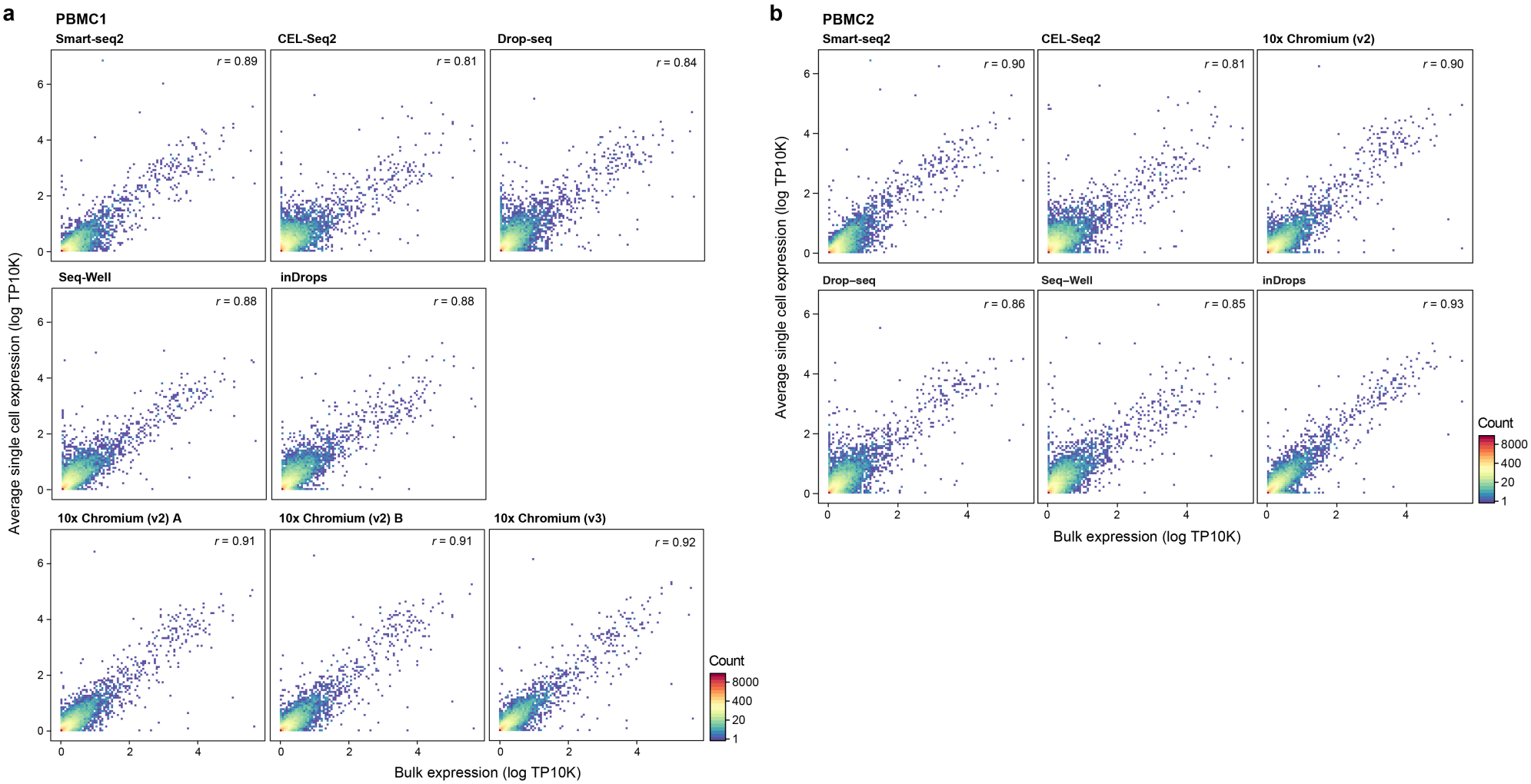
Supplementary Figure 9. **Differential expression of immediate early gene expression across replicates.**
Distribution of immediate early gene expression scores (log(sum(TP10K)), *x*-axis) in mouse cells from each of the two Mixture datasets. For each comparison, we fit a linear regression model on the gene expression score with the number of expressed genes per cell and experiment (Mixture1 vs. Mixture2) as covariates, and calculated the associated *p*-values using a two-sided *t*-test (*n* = 2 biologically independent samples). Violin plot elements are defined as in **Fig. 2**.



Supplementary Figure 10. **Properties of top genes differentially detected in each method.**

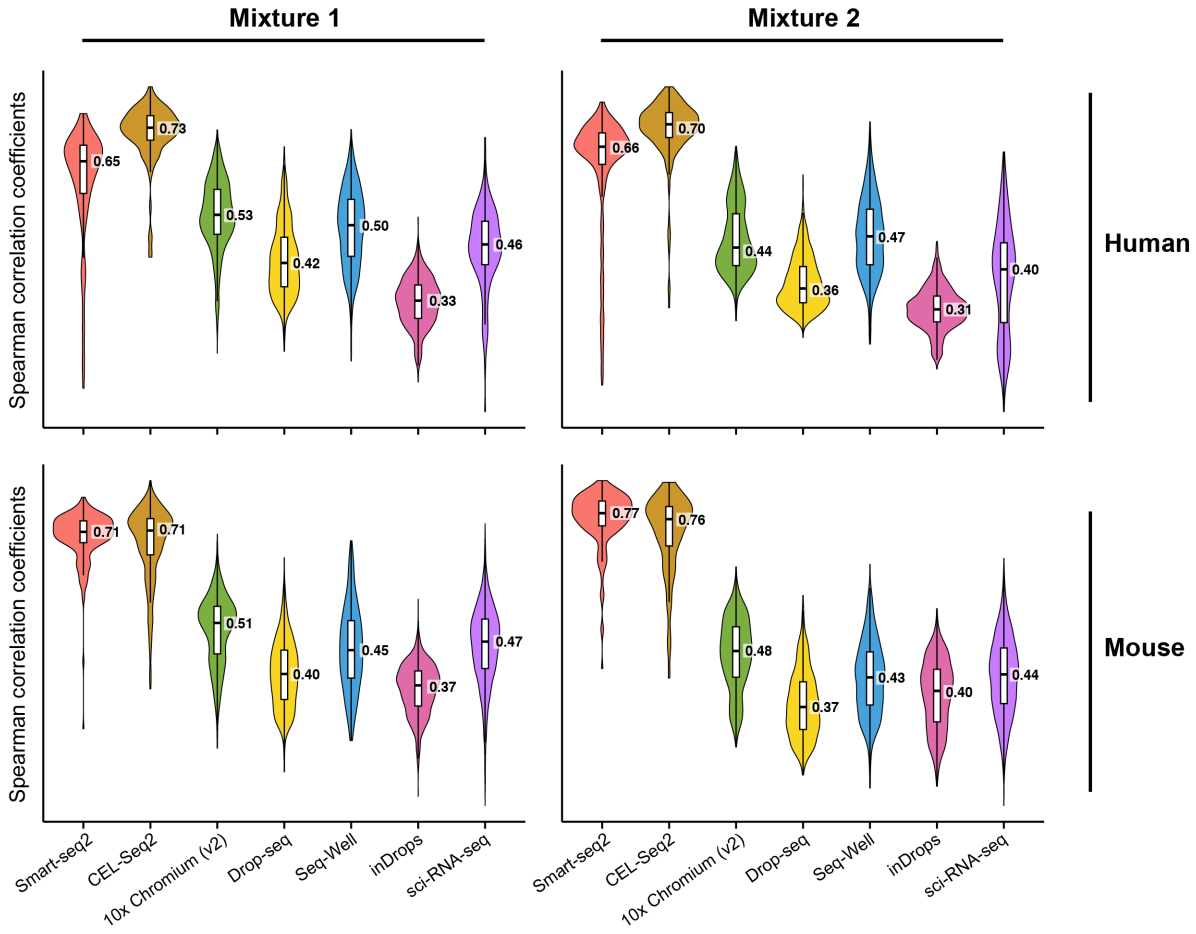
Distribution of GC content (**a**) and length (**b**) for the top 100 genes differentially expressed in each scRNA-seq method relative to all others in the Mixture datasets (*n* = 2 biologically independent samples). Box plot elements are defined as in **Fig. 2**; individual dots are outliers.

Supplementary Figure 11



Supplementary Figure 11. **Correlation of bulk and scRNA-seq PBMC datasets.**

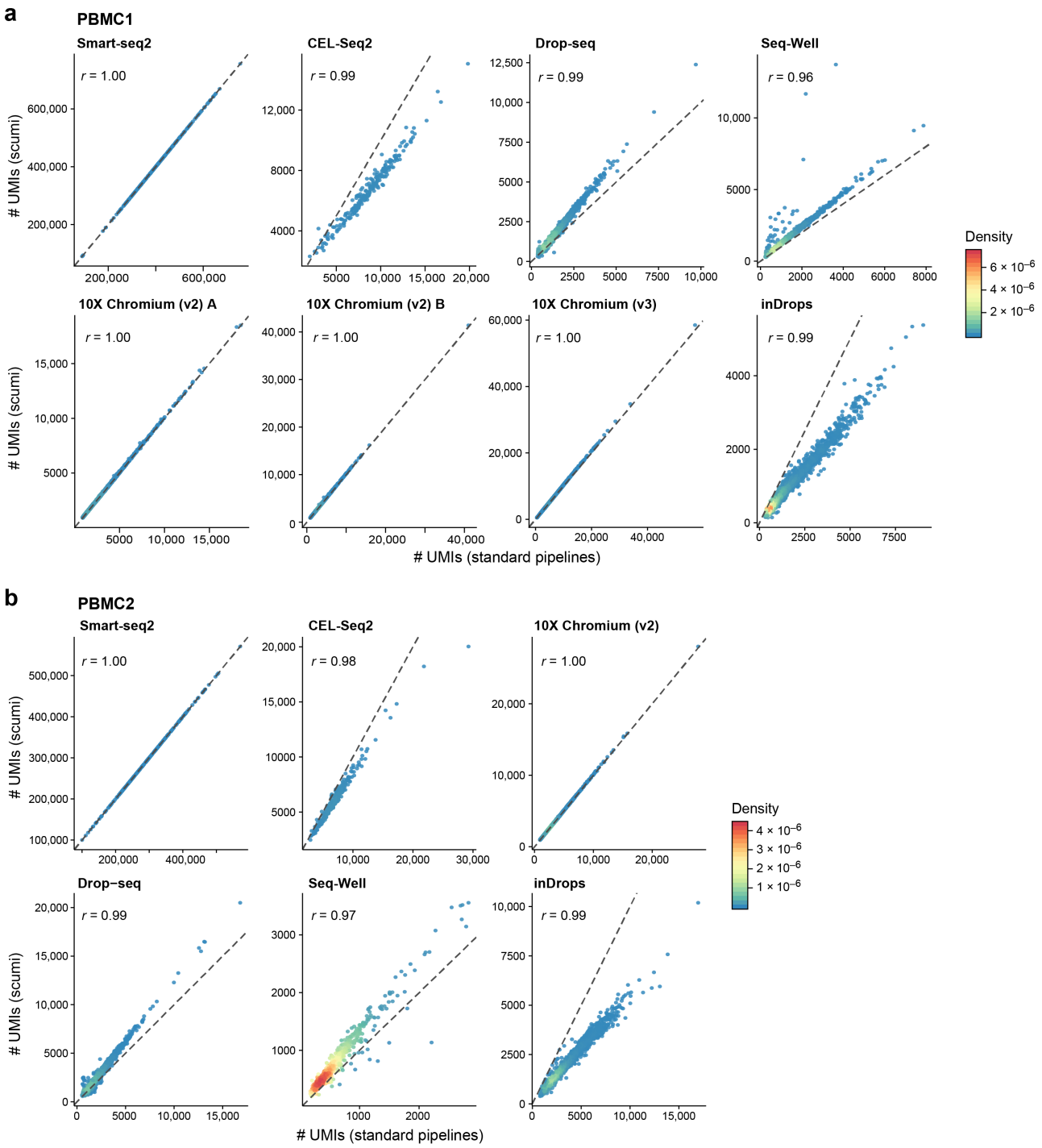
(**a**) PBMC1, (**b**) PBMC2. For each library, we compared the log transcripts per 10K (log TP10K) for bulk and pseudo-bulk from scRNA-seq datasets on a gene by gene basis, plotted as a 2-dimensional histogram (*n* = 1 biologically independent sample per panel). The color of each box in the plane is calculated from the log count of genes in that square. In all cases, the majority of the weight is along the diagonal, suggesting high concordance between bulk and single cell datasets. The Pearson correlation coefficient (*r*) is indicated for each library.

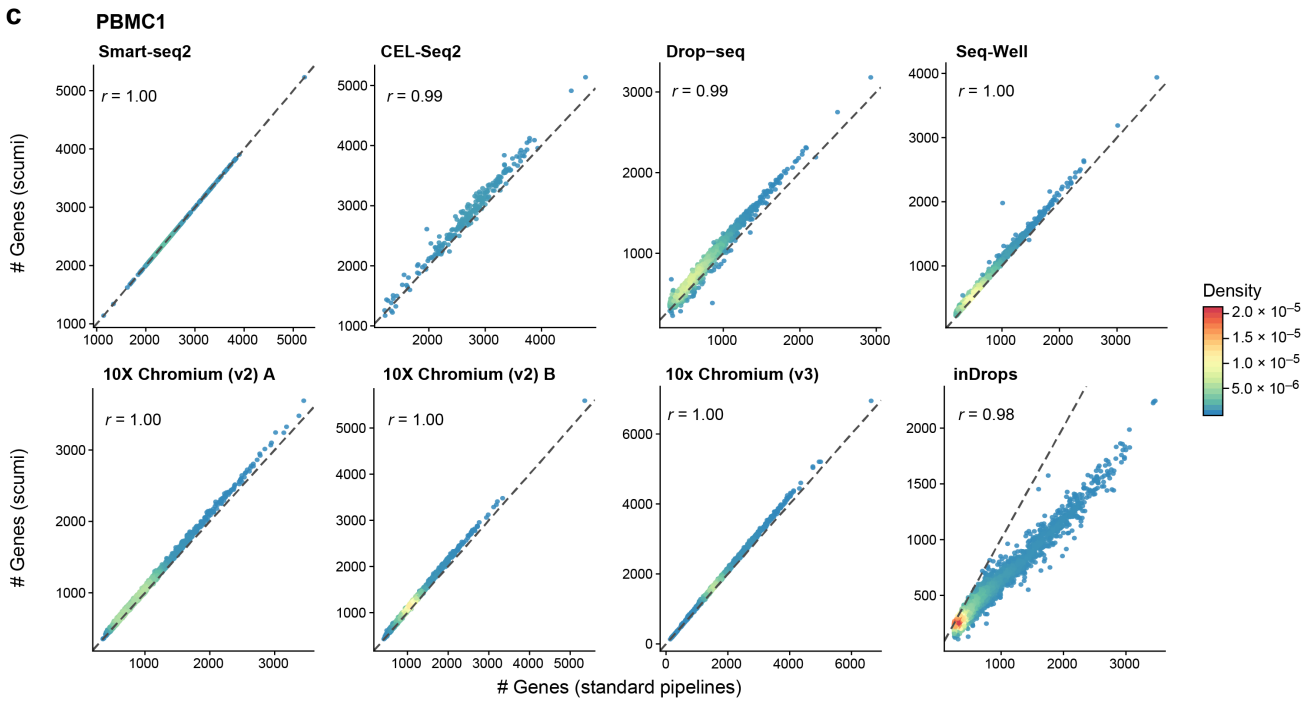


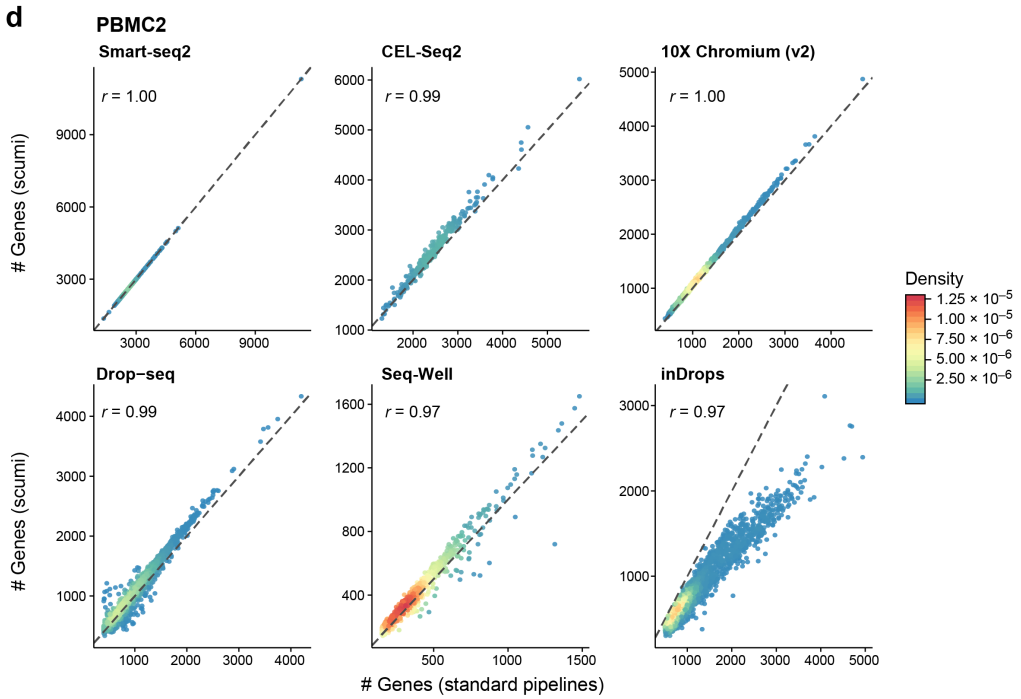
Supplementary Figure 12. **Correlations of single cell and bulk gene expression profiles.**

Distribution of the Spearman correlation coefficients (*y*-axis) between each single cell profile and the corresponding bulk profile for each library (*x*-axis). (*n* = 2 biologically independent samples). Violin and box plot elements are defined as in **Fig. 2**.

Supplementary Figure 13







Supplementary Figure 13. Benchmarking scumi against method-specific pipelines.

(**a**) and (**b**) UMIs per cell, except reads per cell for Smart-seq2. (**c**) and (**d**) genes per cell. Pearson correlation coefficients (*r*) are in the upper left corners. A line indicates the case where the detected number of genes (or UMIs) from both methods are the same. Color encodes the density (two-dimensional kernel density) estimate at each point. (*n* = 1 biologically independent sample per panel).

Supplementary Tables

Supplementary Table 1. Additional reads mapped using transcriptome derived from bulk RNA-seq

Supplementary Table 2. Read sampling.

Supplementary Table 3. Mixture experiment sequencing metrics

Supplementary Table 4. PBMC experiment sequencing metrics

Supplementary Table 5. Cortex experiment sequencing metrics

Supplementary Table 6. PBMC cell types from Harmony not confirmed in separate clustering

Supplementary Table 7. Summary of relative merits of scRNA-seq methods

Supplementary Table 8. Cost and time for each method

Supplementary Table 9. Read processing for each library type

Supplementary Table 10. Gene classifiers for PBMCs

Supplementary Table 11. Gene classifiers for cortex

Supplementary Table 12. Public datasets source and analysis information

Supplementary Table 13. scRNA-seq differentially detected human genes

Supplementary Table 14. scRNA-seq differentially detected mouse genes

Supplementary Table 15. Cell preparation for mixture experiment

Supplementary Table 16. Cell passage information for mixture experiment

Supplementary Table 17. Sequencing runs.

Supplementary Table 18. Adapter sequences used for analysis of unmapped reads

Supplementary Table 19. Optimized clustering parameters for PBMC and cortex datasets