Massive single-cell RNA-seq analysis and imputation via deep learning

Yue Deng[1,*], Feng Bao[2,*], Qionghai Dai[2], Lani F. Wu[1,#], Steven J. Altschuler[1,#]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA.

[2]Department of Automation, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, P. R. China.

*These authors contributed equally to this work.

#To whom correspondence should be addressed:

lani.wu@ucsf.edu, steven.altschuler@ucsf.edu

Recent advances in large-scale RNA-seq enable fine-grained characterization of phenotypically distinct cellular states within heterogeneous tissues. We present scScope, a scalable deep-learning based approach that can accurately and rapidly identify cell-type composition from millions of noisy single-cell gene-expression profiles.

Single-cell RNA-seq (scRNA-seq) can provide high-resolution dissection of complex biological systems, including identification of rare cell subpopulations in heterogeneous tissues, elucidation of cell-developmental programs, and characterization of cellular responses to perturbations[1-5]. Recent platforms, such as DropSeq[6, 7] and GemCode[8], have enabled large-scale scRNA-seq on millions of cells at a time, which offers an unprecedented resolution at which to dissect cell-type compositions.

However, these advances have led to two acute challenges. First, single-cell profiles are highly susceptible to transcript amplification noise and dropout events[9] and such artifacts can become even more pronounced as tradeoffs are made to sequence larger numbers of cells. Second, current analytical packages[9-13] are unable to scale to large datasets due to computational memory and/or speed restrictions. New approaches are needed to extract informative representations from these extremely noisy, massive, high-dimensional scRNA profiles.

To overcome these challenges, we developed scScope, a software package that uses deep learning to extract informative features (low-dimensional representations of gene expression profiles) per cell from massive single-cell data (**Fig. 1a**). A major innovation of scScope is the design of a self-correcting layer. This layer exploits a recurrent network structure to iteratively perform imputations on zero-valued entries of input scRNA-seq data (**Methods**). Importantly, scScope simultaneously conducts cellular-feature learning and dropout imputation in a joint framework and can be trained in parallel on multiple GPUs (**Fig. 1b**). To our knowledge, scScope is the first parallelized deep-learning framework for unsupervised, single-cell data modeling that can deal with both massive and noisy single-cell expression data.

To calibrate the performance of scScope, we initially made use of simulated single-cell expression profile datasets[9], where “ground truth” is known (**Methods**). We compared our results with published methods (ZIFA[9], SIMILR[10], SMAF[11] and PCA as well as with an unpublished Bayesian generative model, scVI[18] (**Methods** and **Supplementary Table 1**).

We began with moderate-sized datasets, containing 10,000 (10K) single-cell profiles. We first investigated the ability to discover a fixed number of subpopulations (k = 50) with RNA-seq dropout rates ranging from ~60% to ~90% (controlled by “decay rate” α ranging from 1 to 0.15, respectively; **Fig. 1c-d, Supplementary Table 2**). scScope performs similarly to other approaches for moderate dropout rates (**Fig. 1c**) and in terms of

imputation error (**Fig. 1d**). However, scScope outperforms all other approaches in finding subpopulations for high dropout rates (**Fig. 1c**). Additionally, the run time of scScope is significantly faster than other single-cell analytical packages (**Fig. 1e**). Second, we varied the number of hidden subpopulations (*k* between 10 to 200) for a fixed, moderate dropout rate (α = 0.5). As expected, clustering accuracy degraded for all approaches with increasing numbers of hidden clusters. Yet, scScope consistently outperformed all tested methods in identifying subpopulations with higher numbers of hidden clusters (**Supplementary Fig. 1**).

We then tested the ability of scScope to analyze massive data sets, containing one million (1M) single-cell profiles. While existing single-cell computational approaches could also be adapted on large-scale dataset by a down-sampling approach[6] (**Methods**), scScope can extract informative features (based on deep learning) and perform graph clustering (**Methods** and **Supplementary Fig. 2**) directly on millions of cells. Compared with down-sampling-based approaches, scScope greatly outperformed the other off-the-shelf single-cell computational packages for discovering cell-type compositions when either all subpopulations are of comparable size (**Fig. 1f**) or contain rare subpopulations (**Methods** and **Fig. 1g**). Importantly, we found that scScope can accurately identify subpopulations comprising as little as 0.1% of the overall cells, which would be missed by down-sampling the original data. Between the two models that can deal with large-scale data, the training speed of scScope is at least 15 times faster than scVI (**Fig. 1h**). Thus, our calibration showed that scScope can efficiently and accurately identify cell subpopulations from datasets with high dropout rates, large numbers of subpopulations, rare cell types and across a wide range of dataset sizes.

We next evaluated scScope on three experimental single-cell RNA datasets. Our goals were to test the ability of scScope to recover original cell-type compositions (retina dataset), to reveal changes in cell-type composition under perturbed conditions (intestinal dataset), and to scale to large datasets to reveal new subpopulations (brain dataset).

The mouse retina dataset[6] contains 44K cells from dissociated mouse retinas. We applied scScope to the 384 most variable genes identified from our preprocessing step and used PhenoGraph[15] to identify subpopulations (**Methods**). The original study identified 39 cell subpopulations using fine-grained manual adjustment and expert visual inspection, which we took as a reference for our comparisons. Compared to other approaches, scScope automatically identified the most similar clustering (number and assignment) to those reported in the original study (**Fig. 2a**, left) and also exhibited the tightest groupings (**Fig. 2a**, right and **Methods**). We annotated the clusters to cell types based on gene markers reported in the original study (**Supplementary Table 3** and **Methods**). (The subpopulation of pericytes, identified by manual adjustment in the original study, was missed by all tested methods.) Overall, clusters identified by scScope showed the most statistically significant enrichment of specific cell-type markers (larger fold-changes) (**Fig.**

2b) and were highly consistent with previous, microscopy validation estimates of cell-type composition and proportion[16] (**Fig. 2c** and **Supplemental Fig. 3**). Thus, scScope performed well at recovering cell-type compositions that are biologically interpretable.

The intestinal dataset[17] contained ~10K cells obtained from mouse intestines with different infection conditions. We applied scScope to the 1,000 most variable genes (**Fig. 2d** and **Methods**). In the original study, enterocytes were identified as a single cluster. Interestingly, scScope further subdivided this cell type into four subpopulations: differential expression of markers provided delineation of distal *vs*. proximal enterocyte subpopulations, while expression levels of these markers provided further delineation into immature *vs*. mature subtypes. The assessment of maturity was based on expression levels of distal or proximal markers (**Fig. 2e** and **Supplementary Fig. 4**). Identifying these refined enterocyte subpopulations allowed us to make predictions about specific cell-type response to infection. For example, the pro-inflammatory gene *Saa1* was overexpressed during both *Salmonella* and *H. polygyrus* (day 10) infections in distal enterocytes, but not in proximal enterocytes (**Fig. 2f** and **Supplementary Table 4**). This geographic pattern of *Saa1* expression is known for *Salmonella* infection, but is a novel prediction arising from scScope analysis for *H. polygyrus* infection. Thus, scScope can be used to rapidly explore scRNA-seq data to predict novel gene function and identify new cell subtypes from perturbed conditions.

The brain dataset contained 1.3M cells obtained from multiple brain regions, including the cortex, hippocampus and ventricular zones, of two embryonic mice. We applied scScope to the 1000 most variable genes (**Methods**). scScope automatically identified 36 clusters, and we assigned each cluster to one of three major cell types based on criteria from the Allen Brain Atlas (http://brain-map.org) (**Fig. 2g**-**h** and **Supplementary Table 5**): Glutamatergic neurons, GABAergic neurons and non-neuronal cells. The proportions of neurons and non-neurons identified by scScope were consistent with cell proportions reported by previous brain study[19] (**Fig. 2g**). Our entire analysis of the 1.3M brain cells required only ~2 hours and 64GB memory (**Fig. 2i**, left and **Supplementary Fig. 5**); in comparison, the original study reported requiring over 350 hours and 300 GB memory to process the same data. Using down-sampling to analyze 1.3M cells with competing single-cell packages revealed fewer cell subpopulations (**Supplementary Fig. 6**), and cell proportions identified by these methods showed poorer agreement with reference proportions (neuron *vs.* non-neuron, **Fig. 2j**). We desired to determine whether we could identify biological meaning to the 36 clusters, some of which contained fewer than 1000 cells. Satisfyingly, by comparing our top overexpressed genes with known cell-type markers[19-21] (**Supplementary Table 6**), we were able to assign two thirds of the clusters to known cell types (**Fig. 2h**). Thus scScope can rapidly, automatically and directly identify *bona fide*, rare cell types from datasets with over a million single-cell transcriptional profiles.

Taken together, scScope offers a platform that will keep pace with the rapid advances in RNA-seq, enabling rapid exploration and accurate dissection of heterogeneous biological states within extremely large datasets of single-cell transcriptional profiles containing dropout measurements.

Acknowledgements

We thank Drs. Jeremy Chang, Satwik Rajaram and Laura Sanman for their helpful comments. We gratefully acknowledge the support of NIH R01 EY028205 and GM112690, NSF PHY-1545915 and SU2C/MSKCC 2015-003 to SJA, NCI-NIH RO1 CA185404 & CA184984 to LFW and the Institute of Computational Health Sciences (ICHS) at UCSF to SJA and LFW.

Author contributions

Y.D., F.B. and Q.D. developed the deep learning algorithms. Y.D. and F.B. conducted experimental analysis on both simulated and biological datasets. The manuscript was written by Y.D., F.B., L.F.W. and S.J.A. All authors read and approved the manuscript.

Competing financial interests

The authors declare no competing interests.

References

1. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* **17**, 175-188 (2016).
2. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610-620 (2015).
3. Saliba, A.-E., Westermann, A.J., Gorski, S.A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research* **42**, 8845-8860 (2014).
4. Shalek, A.K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-+ (2014).
5. Rajaram, S. et al. Sampling strategies to capture single-cell heterogeneity. *Nature Methods* **14**, 967-+ (2017).
6. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
7. Zheng, G.X.Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8** (2017).
8. Eisenstein, M. Startups use short-read data to expand long-read sequencing market. *Nature Biotechnology* **33**, 433-435 (2015).
9. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16** (2015).
10. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods* **14**, 414-+ (2017).
11. Cleary, B., Le, C., Cheung, A., Lander, E.S. & Regev, A. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell* **171**, 1424-+ (2017).
12. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14**, 483-+ (2017).
13. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* (2018).
14. Bengio, Y. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning* **2**, 1-127 (2009).
15. Levine, J.H. et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197 (2015).
16. Jeon, C.J., Strettoi, E. & Masland, R.H. The major cell populations of the mouse retina. *Journal of Neuroscience* **18**, 8936-8946 (1998).
17. Haber, A.L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-+ (2017).
18. Lopez, R., Regier, J., Cole, M.B., Jordan, M. & Yosef, N. Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing. *bioRxiv* (2018).
19. Rosenberg, A.B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-+ (2018).
20. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience* **19**, 335-+ (2016).

21. Mi, D. et al. Early emergence of cortical interneuron diversity in the mouse embryo. *Science* **360**, 81-85 (2018).
22. Franke, L. et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics* **78**, 1011-1025 (2006).
23. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
24. Abadi, M. et al. in USENIX Symposium on Operating Systems Design and Implementation, Vol. 16 265-283 (2016).
25. Hubert, L. & Arabie, P. Comparing Partitions. *Journal of Classification* **2**, 193-218 (1985).
26. Rand, W.M. Objective Criteria for Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846-850 (1971).
27. Calinski, T. A Dendrite Method for Cluster Analysis. *Biometrics* **24**, 207-& (1968).
28. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740-U184 (2014).

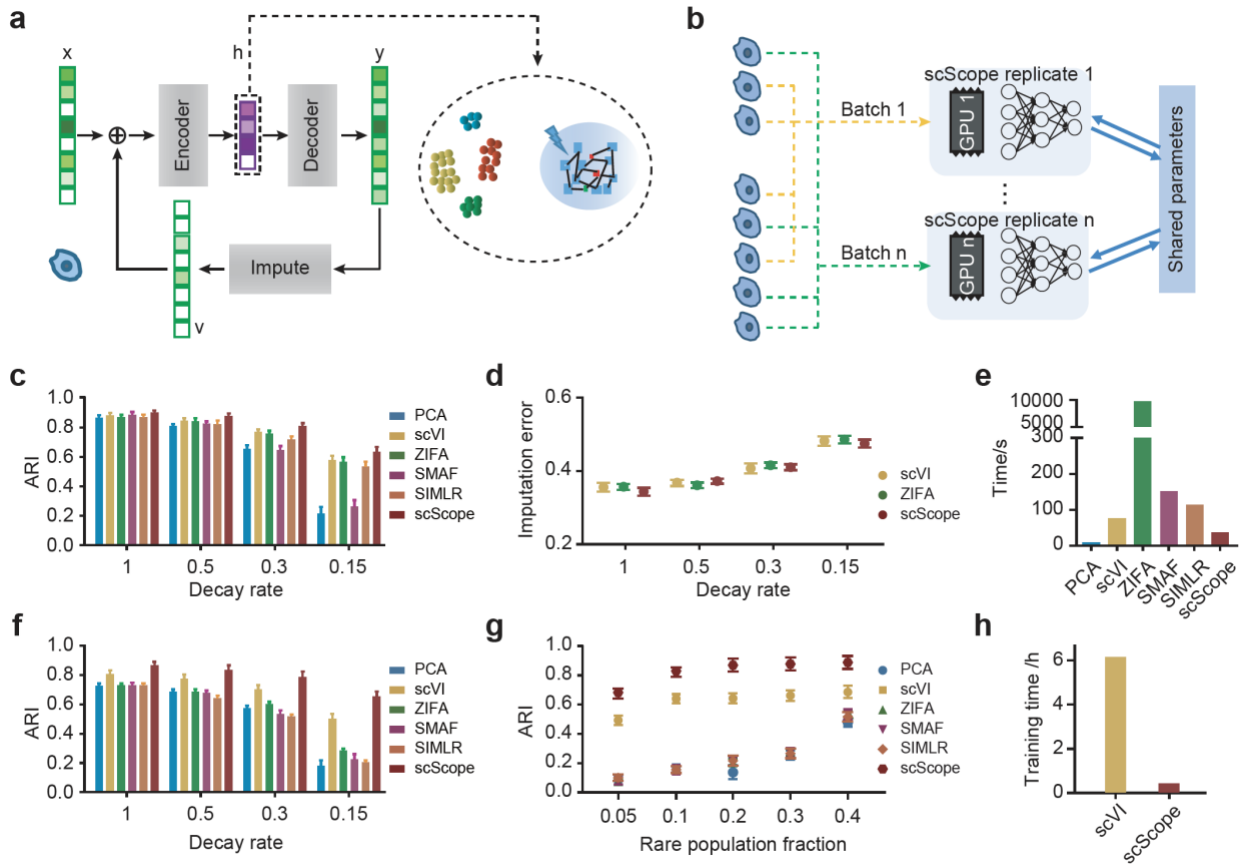

Fig. 1. Overview of scScope architecture and performance on simulated datasets of 10K and 1M single-cell profiles.

a-**b**) **Overview of the recurrent network architecture of scScope.** (**a**) An input single-cell profile x with dropout gene measurements (white entries) is sequentially processed by an encoder layer (for feature extraction), decoder layer (for noise reduction) and imputation layer (for dropout imputation). The imputed vector v is added back to the raw input profile x to fill in missing values. This process proceeds recursively T times to produce a final signature feature vector output h used for biological discovery, such as identification of phenotypically distinct subpopulations (left) or gene network response to perturbations (right). (**b**) Training of scScope is implemented on multiple GPUs to enable fast learning and low memory costs (**Methods**).

c-**e**) **Comparison of clustering, imputation accuracy and run time on simulated single-cell datasets of 10K single-cell profiles.** Method comparisons for varying fractions of dropout genes (controlled by decay rate). Accuracy measurement is based on (**c**) adjusted Rand index or (**d**) fractional measure of imputation errors. (**e**) Comparison of run time performance (decay rate = 0.5). On each simulated condition, 10 random replicates were simulated and tested by different approaches; mean values and standard deviations (error bars) are reported on (**c**-**d**).

f-**h**) **Comparison of clustering accuracy and training time on simulated single-cell datasets of 1M single-cell profiles.** Method comparisons for varying fractions of dropout genes with varying (**f**) fractions of dropout genes or (**g**) rare population fractions. Adjusted Rand index (Mean + STD) was calculated on 10 replicates with the same parameters. scScope and scVI directly analyzed the whole dataset while other methods analyzed 20K down-sampled subset. (**h**) Comparisons of training costs on the full 1M dataset.

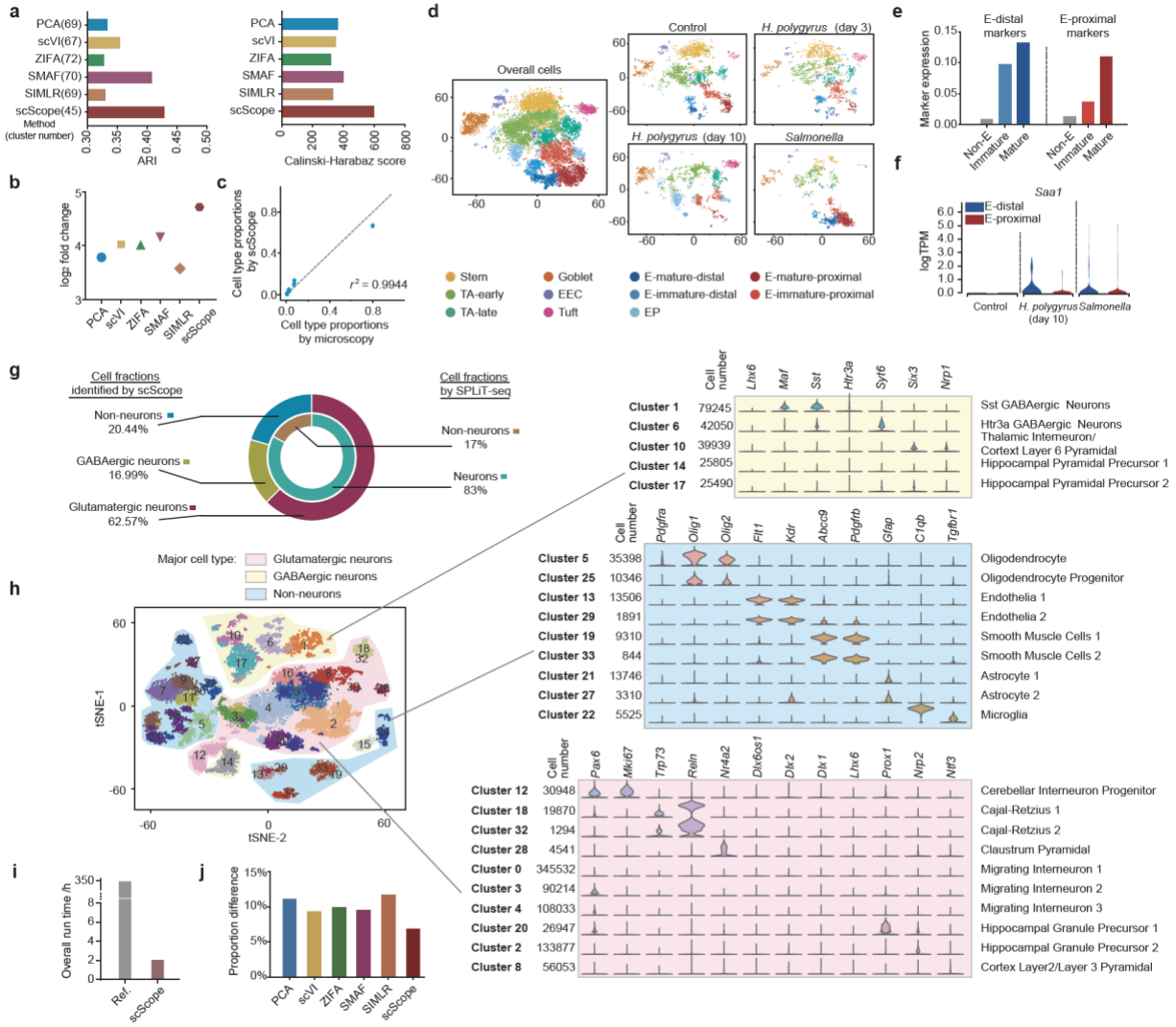

Fig. 2. Application of scScope to identify cell compositions on three experimental mouse scRNA-seq datasets

a-**c**) **Analysis of retina scRNA-seq dataset.** (**a**) Comparison of clustering consistency with original study performed on mouse retina dataset based on adjusted Rand index (left) and Calinski-Harabaz scores (right). Cluster numbers derived from each method are indicated in parentheses (left). (**b**) Averaged fold-change comparisons of cell-type discovery for different methods. (**c**) Consistency between cell proportions identified by scScope and reported cell proportions by microscopy.

d-**f**) **Analysis of intestinal scRNA-seq dataset.** (**d**) Changes in cell-type composition of mouse intestinal epithelial under different infection conditions, visualized via tSNE plots. TA: transit amplifying, EEC: enteroendocrine, EP: enterocyte progenitor, E: enterocyte. scScope identified four subtypes of enterocyte cells. (**e**) Identification of mature vs. immature and distal vs. proximal enterocyte subpopulations. Shown are expression levels

of E-distal and E-proximal gene markers (average UMI count) on the four enterocyte subtypes predicted by scScope and, for comparison, all other clusters (Non-E). (**f**) Discovery of differential expression of the gene *Saa1* in distal and proximal enterocytes after *Salmonella* and *H. polygyrus* infections.

g-**j**) **Analysis of brain scRNA-seq dataset.** (**g**) Fractions of three major cell types (Glutamatergic neurons, GABAergic neurons and non-neurons) identified by scScope and comparisons with reported neuron fractions by previous SPLiT-seq research. (**h**) (**left**) scScope results visualized using tSNE on 30K cells (randomly sampled from the full dataset). Clusters were divided to three major types based on gene markers. (**right**) Large-scale annotation of clusters to known cell types according to top 10 overexpressed genes. Vertical axis (left): clusters with known cell type annotations and corresponding cell numbers. Horizontal axis: differentially expressed marker genes across shown clusters. Vertical axis (right): previously reported cell subtype-specific genes, used to provide cluster annotation. (**i**) Comparisons of overall run time between scScope, and previously reported reference ("Ref.") results (left) on 1.3 million mouse brain single-cell profiles. (**j**) Differences between neuron fractions obtained by comparison methods and known proportions on 1.3M mouse brain data. scScope and scVI were directly applied to the whole dataset while other methods required down-sampling to 20K before performing analysis.