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.

However, these advances have led to two acute challenges. First, single-cell profiles are highly susceptible to transcript amplification noise and dropout events[9]. Second, current analytical packages[9-13] are unable to conduct learning on large-scale 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. To our knowledge, scScope is the first deep-learning framework for unsupervised, single-cell data modeling that can deal with both massive and noisy single-cell expression data.

We first evaluated the performance of scScope on simulated single-cell expression profile datasets[9], where “ground truth” is known. A 10K-sized dataset was used for extensive comparisons of methods (ZIFA[9], SIMILR[10], SMAF[11], PCA and Auto Encoder (AE)[14]; **Methods**); a 1M-sized dataset was used to compare scalability of scScope but, due to performance limitations of current off-the-shelf software packages, could only be compared with two generic machine-learning approaches (PCA and AE). First, we investigated the ability to discover subpopulations with dropout rates ranging from ~40% to ~90% (decay rate α ranging from 5 to 0.15, respectively; **Methods** and **Supplementary Table 1**) with a fixed number (k = 50) of cell subpopulations (**Fig. 1b** top and **Methods**). We found scScope performs similarly to other approaches for moderate dropout rates, but outperforms all others as dropout rates increase. Second, we compared imputation accuracy of scScope (**Methods**) with AE and ZIFA, two other approaches able to impute missing data (**Fig. 1b** bottom). On the 10K dataset, scScope and ZIFA perform similarly

in imputation accuracy and outperform AE, but the running time of scScope is ~100X times faster than ZIFA (**Supplementary Table 2**). Only scScope and AE can scale to perform feature learning directly on the 1M dataset, where scScope performs significantly better than AE. Additionally, we varied the number of hidden subpopulation numbers for a fixed, moderate dropout rate (α = 0.5). All approaches degraded in accuracy with increasing numbers of hidden clusters, however scScope consistently outperformed all tested methods (**Supplementary Fig. 1**). Finally, we tested our ability to identify rare subpopulations within the 1M dataset. We automatically determined subpopulation numbers; for this, on the 10K dataset we used PhenoGraph[15] and for the 1M dataset we developed a scalable graph clustering approach that integrated PhenoGraph (**Methods** and **Supplementary Fig. 2**). Our results showed that we can accurately identify subpopulations comprising as little as 0.1% of the overall cells, which would be missed by down-sampling the original data (**Fig. 1c, Methods**). Thus, scScope provides a scalable and accurate approach for imputing missing data in complicated conditions (*e.g.* sparse measurements with large hidden population number) and identifying rare subpopulations within massive datasets.

We next applied scScope to three different single-cell RNA datasets covering a wide range of sample sizes, from 10^4 – 10^6 cells. As preprocessing steps for all datasets, we log-transformed gene expressions (*i.e.* ln(TPM + 1)) and identified the genes that varied the most from cell to cell for subsequent analysis (**Methods**). As we lack ground truth for these experimental datasets, we assessed performance based on biological interpretability and comparison to previously published results.

To test scalability, we first analyzed a scRNA-seq dataset containing 1.3M mouse brain cells (**Methods**) obtained from multiple brain regions, including the cortex, hippocampus and ventricular zones, of two embryonic mice. As above, we made use of our scalable clustering approach to identify subpopulations. Our entire analysis of the 1.3M brain cells required only ~2 hours and 64GB memory (**Fig. 1d** and **Supplementary Fig. 3**); in comparison, the original study reported requiring over 350 hours and 300 GB memory to process the same data. To give biological meaning to the 36 clusters automatically identified by scScope, we assigned each to one of three major cell types based on criteria from the Allen Brain Atlas (http://brain-map.org) (**Fig. 1e** and **Supplementary Table 3**): excitatory neurons (Glutamatergic), inhibitory neurons (GABAergic) and non-neuronal cells. By comparing our top overexpressed genes ($\log_2$ fold-change > 3) with known cell-type markers[16], we were able to assign a subset of scScope clusters to: neurogliaform cells (*Ndnf*, *Reln*), L6a corticothalamic glutamatergic neurons (*Nts*), *Htr3a*$^+$ *Vip*$^-$ GABAergic neurons, oligodendrocytes (*Olig1, Olig2*), microglial cells (*C1qb*) and endothelial cells (*Flt1*) (**Fig. 1f**). Thus, scScope allowed us to rapidly explore cellular heterogeneity within more than a million single-cell transcriptional profiles.

To test accuracy, we next compared scScope with other approaches on a dataset containing ~44K retinal cells sequenced by DropSeq[6], using 384 highly variable genes identified from our preprocessing step and 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** top) and also exhibited the tightest groupings (**Fig. 2a** bottom and **Methods**). To interpret biological meaning, we annotated the clusters to one of 12 cell types based on gene markers reported in the original study (**Supplementary Table 4** and **Methods**). (The subpopulation of pericytes, identified by manual adjustment in the original study, was missed by all tested methods.) scScope slightly outperformed all other tested methods in dissecting cell-type composition[17] (**Supplemental Fig. 4**). Further, the annotations of cell clusters identified by scScope were, for the overall average, the most statistically significant (larger fold-change and smaller rank-sum P-values) (**Fig. 2b**, **Supplementary Fig. 5** and **Supplementary Table 5**). Thus, scScope is able to identify meaningful, fine-grained cellular subtypes.

Finally, to evaluate the ability of scScope to identify changes in cell-type composition and transcriptional state under different conditions, we made use of a DropSeq dataset of mouse intestinal epithelial exposed to different pathogen infections[18] (**Methods**). We applied scScope to the pooled set of 9,842 cells, using the 1000 most highly variable genes (**Fig. 2c**, **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. 2d** and **Supplementary Fig. 6**). 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. 2e** and **Supplementary Table 6**). This geographic pattern of *Saa1* expression is known for *Salmonella* infection, but is a novel prediction arising from scScope analysis for *H. polygyrus* infection.

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 Al. *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. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience* **19**, 335-+ (2016).
17. Jeon, C.J., Strettoi, E. & Masland, R.H. The major cell populations of the mouse retina. *Journal of Neuroscience* **18**, 8936-8946 (1998).
18. Haber, A.L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-+ (2017).
19. 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).
20. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

21. Abadi, M. et al. in USENIX Symposium on Operating Systems Design and Implementation, Vol. 16 265-283 (2016).
22. Hubert, L. & Arabie, P. Comparing Partitions. *Journal of Classification* **2**, 193-218 (1985).
23. Rand, W.M. Objective Criteria for Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846-850 (1971).
24. Calinski, T. A Dendrite Method for Cluster Analysis. *Biometrics* **24**, 207-& (1968).
25. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740-U184 (2014).

Legends

Fig. 1. Overview of scScope architecture and performance on large-scale datasets.

a) Overview of the recurrent network architecture of scScope. 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-**c**) Comparison of clustering and imputation accuracy on simulated single-cell datasets with varying fractions of dropout genes (controlled by decay rate) and sizes, as well as ability to identify rare subpopulations. Accuracy measurement is based on Adjusted Rand index (**b** top and **c**) or fractional measure of imputation errors (**b** bottom). On each simulated condition, 10 random replicates were simulated and tested by different approaches; mean values and standard deviations (error bars) are reported.

e) Comparisons of run time and memory requirements between scScope and previous study on a dataset containing 1.3 million mouse brain single-cell profiles.

f) 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 (legend).

g) Assignment of clusters to cell types. Vertical axis (left): clusters with genes differentially expressed ($\log_2$ fold-change > 3) compared to all other clusters. Horizontal axis: differentially genes across shown clusters. Red boxes: previously reported cell subtype-specific genes, used to provide cluster annotation (right).

Fig. 2. Application of scScope to identify changes in cell-type composition and transcriptional state.

a) Comparison of clustering consistency with original study performed on mouse retina dataset based on Adjusted Rand index (ARI, top) and Calinski-Harabaz scores (bottom). Cluster numbers derived from each method are indicated in parentheses (top).

b) Comparisons of cell-type discovery on mouse retina dataset for different methods. Expression differences for cell markers were evaluated by averaged $\log_2$ fold-changes on each cell type. Right-most panel: averaged fold-changes across different cell types.

c) 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.

d) 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).

e) Differential expression of the gene *Saa1* in distal and proximal enterocytes after *Salmonella* and *H. polygyrus* infections.

Figures

Figure 1

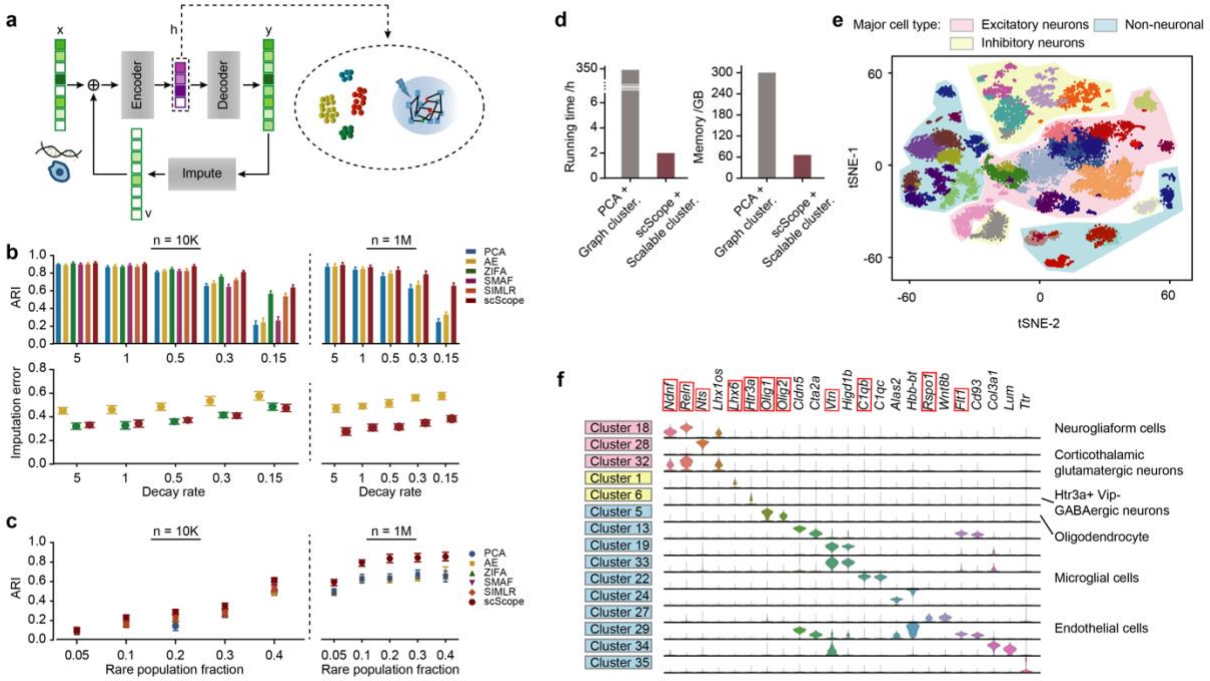
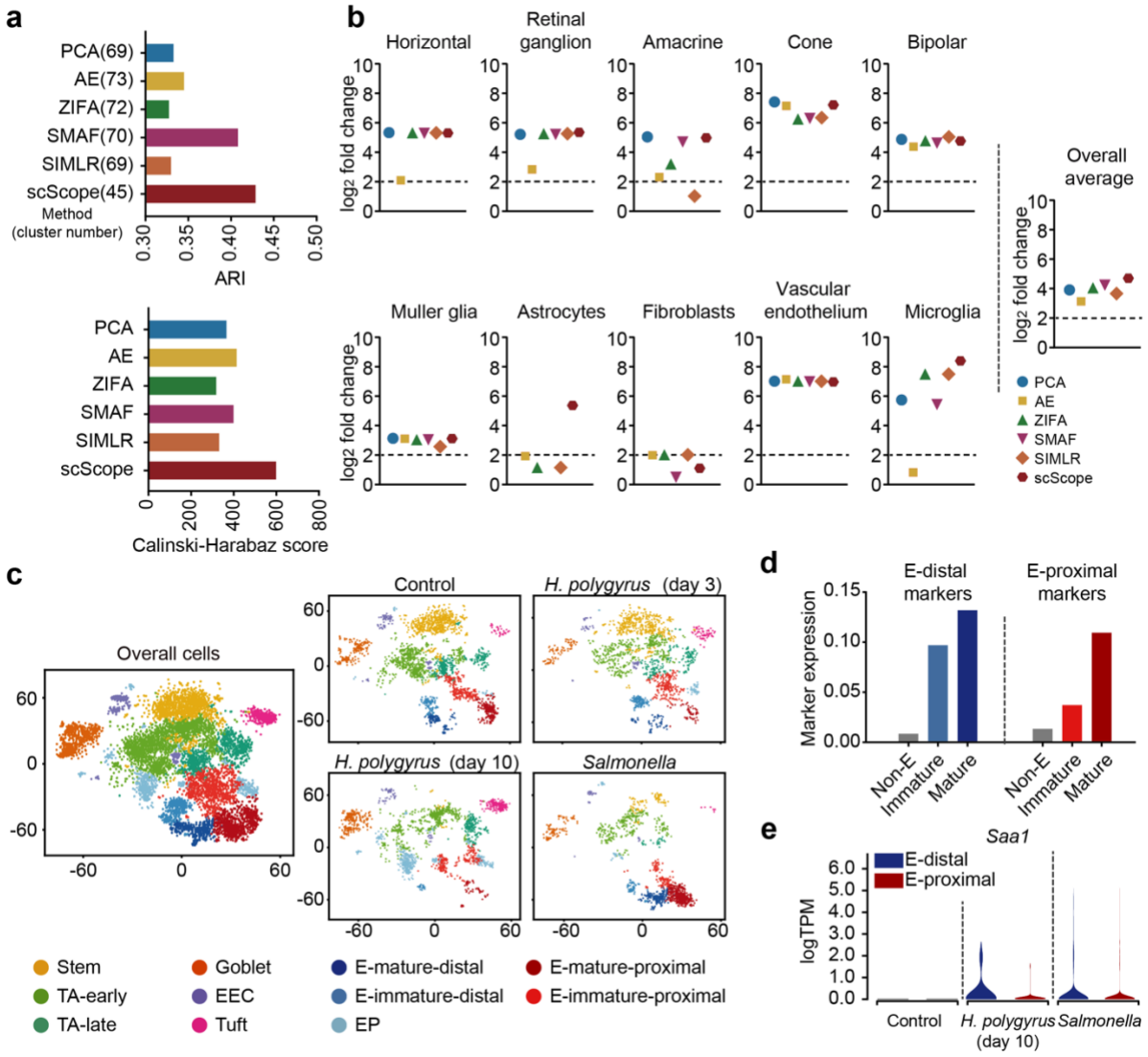


Figure 2


a
PCA(69)
AE(73)
ZIFA(72)
SMAF(70)
SIMLR(69)
scScope(45)
Method
(cluster number)
0.30
0.35
0.40
0.45
0.50
ARI
PCA
AE
ZIFA
SMAF
SIMLR
scScope
0
200
400
600
800
Calinski-Harabaz score
b
Horizontal
Retinal ganglion
Amacrine
Cone
Bipolar
Muller glia
Astrocytes
Fibroblasts
Vascular endothelium
Microglia
Overall average
log2 fold change
PCA
AE
ZIFA
SMAF
SIMLR
scScope
c
Overall cells
Control
H. polygyrus (day 3)
H. polygyrus (day 10)
Salmonella
Stem
TA-early
TA-late
Goblet
EEC
Tuft
E-mature-distal
E-immature-distal
EP
E-mature-proximal
E-immature-proximal
d
E-distal markers
E-proximal markers
Marker expression
Non-E
Immature
Mature
e
Saa1
E-distal
E-proximal
logTPM
Control
H. polygyrus (day 10)
Salmonella