

1   **Deep learning enables accurate clustering and batch effect removal in single-cell  
2 RNA-seq analysis**

3  
4   Xiangjie Li<sup>1,2</sup>, Yafei Lyu<sup>1</sup>, Jihwan Park<sup>3</sup>, Jingxiao Zhang<sup>2</sup>, Dwight Stambolian<sup>4</sup>, Katalin Susztak<sup>3</sup>,  
5   Gang Hu<sup>1,5\*</sup>, Mingyao Li<sup>1\*</sup>

6   1) Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania  
7   Perelman School of Medicine, Philadelphia, PA 19104, USA.

8   2) Center for Applied Statistics, School of Statistics, Renmin University, Beijing, China.

9   3) Departments of Medicine and Genetics, University of Pennsylvania Perelman School of  
10   Medicine, Philadelphia, PA 19104, USA.

11   4) Department of Ophthalmology, University of Pennsylvania Perelman School of Medicine,  
12   Philadelphia, PA 19104, USA.

13   5) Department of Information Theory and Data Science, School of Mathematical Sciences and  
14   LPMC, Nankai University, Tianjin 300071, China.

15  
16   \*Correspondence:

17  
18   Gang Hu, Ph.D.

19   [huggs@nankai.edu.cn](mailto:huggs@nankai.edu.cn)

20  
21   Mingyao Li, Ph.D.

22   [mingyao@pennmedicine.upenn.edu](mailto:mingyao@pennmedicine.upenn.edu)

23  
24  
25   **Single-cell RNA sequencing (scRNA-seq) can characterize cell types and states through  
26   unsupervised clustering, but the ever increasing number of cells imposes computational  
27   challenges. We present an unsupervised deep embedding algorithm for single-cell  
28   clustering (DESC) that iteratively learns cluster-specific gene expression signatures and  
29   cluster assignment. DESC significantly improves clustering accuracy across various  
30   datasets and is capable of removing complex batch effects while maintaining true  
31   biological variations.**

32  
33   A primary challenge in scRNA-seq analysis is analyzing the ever increasing number of cells,  
34   which can be thousands to millions in large projects such as the Human Cell Atlas<sup>1</sup>. Identifying  
35   cell populations is a challenge in large datasets because many existing scRNA-seq clustering  
36   methods cannot be scaled up to handle them. It is desirable to first learn cluster-specific gene  
37   expression features from cells that are easy to cluster because they provide valuable information  
38   on cluster-specific gene expression signatures. These cells can help improve clustering of cells  
39   that are hard-to-cluster.

40  
41   Another challenge in scRNA-seq analysis is batch effect, which is systematic gene expression  
42   difference from one batch to another<sup>2</sup>. Batch effect is inevitable in studies involving human tissues  
43   because the data are often generated at different times and the batches can confound biological  
44   variations. Failure to remove batch effect will complicate downstream analysis and leads to a false  
45   interpretation of results.

46  
47   ScRNA-seq clustering and batch effect removal are typically addressed through separate  
48   analyses. Commonly used approaches to remove batch effect include Seurat's Canonical

49 Correlation Analysis<sup>3</sup> (CCA) or Mutual Nearest Neighbors (MNN) approach<sup>4</sup>. After removing batch  
50 effect, clustering analysis is performed to identify cell clusters using methods such as Louvain's  
51 method<sup>5</sup>, Infomap<sup>6</sup>, graph-based clustering<sup>7</sup>, shared nearest neighbor<sup>8</sup>, or consensus clustering  
52 with SC3<sup>9</sup>. Since some cell types are more vulnerable to batch effect than others, batch effect  
53 removal should be performed jointly with clustering to achieve optimal performance. However,  
54 none of the existing methods are capable of simultaneously clustering cells and removing batch  
55 effect.

56  
57 We developed DESC, an unsupervised deep learning algorithm that iteratively learns cluster-  
58 specific gene expression representation and cluster assignments for scRNA-seq data clustering  
59 (**Fig. 1a**). Using a deep neural network, DESC initializes clustering obtained from an autoencoder  
60 and learns a non-linear mapping function from the original scRNA-seq data space to a low-  
61 dimensional feature space by iteratively optimizing a clustering objective function. This iterative  
62 procedure moves each cell to its nearest cluster, balances biological and technical differences  
63 between clusters, and reduces the influence of batch effect. DESC also enables soft clustering  
64 by assigning cluster-specific probabilities to each cell, facilitating the clustering of cells with high-  
65 confidence.

66  
67 We benchmarked DESC's performance by analyzing the multi-tissue gene expression data in  
68 GTEx<sup>10</sup>. We treat this dataset as the gold-standard because the tissue origins are known.  
69 Although not generated by scRNA-seq, GTEx data are similar to scRNA-seq in that it contains a  
70 large number of samples (n=11,688) originated from many tissue types and is similar to the  
71 volume and complexity of scRNA-seq data (**Supplementary Note 1**). DESC's clustering yields  
72 an adjusted rand index (ARI) of 0.790, whereas the ARIs for Louvain's method, SC3, and Infomap  
73 are 0.755, 0.349, and 0.267, respectively. As shown in the Sankey diagrams (**Supplementary**  
74 **Fig. 3**), samples that were misclassified by DESC and Louvain's method tend to be from closely  
75 related tissues, whereas SC3 tends to misclassify samples from tissues distantly related.

76  
77 We analyzed a scRNA-seq dataset generated from the midbrain of Drosophila, which includes  
78 10,286 cells using Drop-seq<sup>11</sup> (**Supplementary Note 2**). This dataset has minimal batch effect.  
79 DESC identified three types of mushroom body Kenyons, with 1,038 out of the 1,053 Kenyon cells  
80 correctly classified, a 98.6% classification accuracy (**Fig. 1b**). DESC also separated cholinergic,  
81 glutamatergic, and GABAergic neurons, which were mixed together in the Louvain's clustering as  
82 shown in the original paper (**Fig. 1c**). These results indicate that DESC can identify cell types that  
83 are detectable by the Louvain's method, and is also able to separate more closely related cells,  
84 indicating its increased accuracy in classifying closely related cell types. We further applied  
85 Louvain's method to the low-dimensional representation learned from the autoencoder in DESC  
86 for clustering, and separated the cholinergic, glutamatergic, and GABAergic neurons better than  
87 the original Louvain's clustering with principal components (PC) based dimension reduction  
88 (**Supplementary Fig. 4**). These results suggest the autoencoder is more effective than PC in  
89 dimension reduction for single-cell clustering.

90  
91 Encouraged by these findings, we analyzed a scRNA-seq dataset with known batch effect  
92 (**Supplementary Note 3**). Shekhar et al.<sup>12</sup> sequenced 23,494 retinal bipolar cells using Drop-seq,  
93 where cells from six replicates were processed in two different batches. **Fig. 2a and 2c** show that  
94 DESC removed the batch effect, and yields an ARI of 0.973 for clustering. The corresponding  
95 ARIs for Louvain, SC3, Infomap, CCA and MNN are 0.965, 0.521, 0.560, 0.637, and 0.974,  
96 respectively. Although DESC, Louvain, and MNN have similar ARIs, DESC has the smallest  
97 Kullback-Leibler (KL) divergence, which measures the degree of random mixing of cells in  
98 different batches, indicating that DESC is more effective in removing batch effect (**Fig. 2b**).  
99 Further analysis revealed that the batch effect removal in DESC is due to its iterative clustering,

100 in which cells from the same cluster, separated by technical batch effect, are grouped closer and  
101 closer to the cluster centroid over iterations (**Figs. 2d and 2e**).  
102

103 We also assessed the performance of DESC on data with complex batch effect generated from  
104 multiple subjects using the same platform but in different labs. Such complex batch effect is  
105 common in human studies because logistical constraints mandate that data from different  
106 subjects be generated at different times and perhaps in different labs, which result in complex  
107 batch effects that are challenging to address. To examine the robustness of DESC in the presence  
108 of this complex batch effect, we analyzed scRNA-seq data obtained from seven human kidneys  
109 (**Supplementary Note 4**). This dataset includes 8,544 cells, derived from four healthy kidneys,  
110 generated by us using 10X, and 7,149 cells obtained from the normal part of kidneys in three  
111 patients with kidney tumor<sup>13</sup>, also generated by 10X, but in a different lab. **Figs. 3a and b** show  
112 that DESC removed batch effect, with the seven biological samples and the two different datasets  
113 randomly mixed. The KL-divergence is lower for DESC than for CCA and MNN (**Fig. 3d**),  
114 indicating that DESC is more effective in removing batch effect both at the subject level and  
115 dataset level.  
116

117 The kidneys and the immune system are closely linked. It has been shown that the accumulation  
118 of natural killer (NK) cells promotes chronic kidney inflammation and contributes to kidney  
119 fibrosis<sup>14</sup>. T cells, which have a well-described role in renal injury, are involved in renal fibrosis<sup>15</sup>.  
120 Previous studies have shown that NK cells play a role in the regulation of the adaptive immune  
121 response and stimulate or inhibit T cell responses<sup>16</sup>. Better understanding of how different  
122 components of the immune system mediate kidney disease requires a clear separation of NK and  
123 T cells. **Fig. 3c and Supplementary Fig. 8** show that both DESC and MNN identified T cells and  
124 NK cells as separate clusters; however, CCA mixed some of the NK cells with T cells, possibly  
125 due to overcorrection of true biological variations. These results indicate that DESC not only  
126 removed technical batch effect more effectively than CCA and MNN, but also maintained true  
127 biological variations among closely related immune cells.  
128

129 To further demonstrate that DESC preserves true biological variations, we considered an even  
130 more complex situation in which technical batches were completely confounded with biological  
131 variations. This is inevitable in disease studies where tissues are processed immediately to  
132 maintain cell viability resulting in the preparation of normal and diseased samples in different  
133 batches. For data generated in such complex settings, it is desirable to remove technical batch  
134 effect while maintaining true biological variations between normal and diseased samples so that  
135 disease specific subpopulations can be identified. We analyzed a dataset generated by 10X that  
136 includes 24,679 human PBMCs from eight patients with lupus<sup>17</sup> (**Supplementary Note 5**). The  
137 cells were split into a control group and a matched group stimulated with INF-β, which leads to a  
138 drastic but highly cell type-specific response. This dataset is extremely challenging because  
139 removal of technical batch effect is complicated by the presence of biological differences, both  
140 between cell types under the same condition and between different conditions.  
141

142 **Fig. 3c** shows that DESC randomly mixed cells between the control and the stimulus conditions  
143 for all cell types except CD14<sup>+</sup> monocytes. Differential expression (DE) analysis revealed a drastic  
144 change in gene expression after INF-β stimulation for CD14<sup>+</sup> monocytes (**Fig. 3d**); the number of  
145 DE genes and the magnitude of DE, measured by p-value and fold-change, are several orders  
146 more pronounced than the other cell types. This is consistent with previous studies showing  
147 CD14<sup>+</sup> monocytes with a more drastic gene expression change than B cells, dendritic cells, and  
148 T cells after INF-β stimulation<sup>18, 19</sup>. These results suggest that DESC is able to remove technical  
149 batch effect and maintain true biological variations induced by INF-β. MNN also preserved the  
150 biological difference between the control and the INF-β stimulated CD14<sup>+</sup> monocytes, but the NK

151 cells are less well separated from CD8 T cells (**Supplementary Fig. 15a**). CCA masked the  
152 biological difference between the control and the INF- $\beta$  stimulated CD14 $^{+}$  monocytes indicating  
153 that it might have overcorrected batch effect (**Supplementary Fig. 15a**).  
154

155 In summary, we have developed a deep learning algorithm that clusters scRNA-seq data by  
156 iteratively optimizing a clustering objective function with a self-training target distribution. DESC's  
157 memory usage and running time increase linearly with the number of cells, thus making it scalable  
158 to large datasets (**Fig. 3e**). DESC can further speed up computation by GPUs. We analyzed a  
159 mouse brain dataset with 1.3 million cells generated by 10X, which only took about 3.5 hours with  
160 one NVIDIA TITAN Xp GPU (**Supplementary Note 6**). Compared to existing scRNA-seq  
161 clustering methods DESC improves clustering by iteratively learning cluster-specific gene  
162 expression features from cells clustered with high confidence. This iterative clustering also  
163 removes batch effect and maintains true biological differences between clusters. As the growth  
164 of single-cell studies increases, DESC will be a more precise tool for clustering of large datasets.  
165  
166

## 167 ACKNOWLEDGEMENTS

168 This work was supported by the following funding: NIH R01GM108600, R01GM125301, and  
169 R01HL113147 (to M.L.); NIH R01DK076077 (to K.S.).  
170

## 171 AUTHOR CONTRIBUTIONS

172 This study was conceived of and led by M.L. and H.G.. X.L., H.G., and M.L. designed the model  
173 and algorithm. X.L. implemented the DESC software and led the data analysis with input from  
174 M.L., H.G., and J.Z.. Y.L. helped with software development and testing. K.S. and J.P. generated  
175 the human kidney scRNA-seq data, and provided input on the kidney data analysis. D.S. provided  
176 input on the mouse retina scRNA-seq data analysis. M.L., X.L., and H.G. wrote the paper with  
177 feedback from Y.L., J.P., J.Z., D.S., and K.S..  
178

## 179 COMPETING FINANCIAL INTERESETS STATEMENT

180 The authors declare no competing interests.  
181  
182

## 183 REFERENCES

- 185 1. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
- 186 2. Hicks, S.C., Townes, F.W., Teng, M. & Irizarry, R.A. Missing data and technical  
187 variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**(4):562-578  
188 (2017).
- 189 3. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell  
190 transcriptomic data across different conditions, technologies, and species. *Nat  
191 Biotechnol* **36**, 411-420 (2018).
- 192 4. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell  
193 RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat  
194 Biotechnol* **36**, 421-427 (2018).
- 195 5. Blondel, V.D., Guillaume, J.L., Lambiotte, R. & Lefebvre, E. Fast unfolding of  
196 communities in large networks. *J Stat Mech*, 10008-10012 (2008).
- 197 6. Rosvall, M. & Bergstrom, C.T. Maps of random walks on complex networks reveal  
198 community structure. *Proc Natl Acad Sci U S A* **105**, 11118-1123 (2008).
- 199 7. Levine, J.H. et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like  
200 Cells that Correlate with Prognosis. *Cell* **162**, 184-197 (2015).

- 201 8. Xu, C. & Su, Z.C. Identification of cell types from single-cell transcriptomes using a novel  
202 clustering method. *Bioinformatics* **31**, 1974-1980 (2015).
- 203 9. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*  
204 **14**, 483-486 (2017).
- 205 10. Consortium, G.T. et al. Genetic effects on gene expression across human tissues.  
206 *Nature* **550**, 204-213 (2017).
- 207 11. Croset, V., Treiber, C.D. & Waddell, S. Cellular diversity in the Drosophila midbrain  
208 revealed by single-cell transcriptomics. *Elife* **7** (2018).
- 209 12. Shekhar, K. et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-  
210 Cell Transcriptomics. *Cell* **166**, 1308-1323 (2016).
- 211 13. Young, M.D. et al. Single-cell transcriptomes from human kidneys reveal the cellular  
212 identity of renal tumors. *Science* **361**, 594-599 (2018).
- 213 14. Law, B.M.P. et al. Interferon-gamma production by tubulointerstitial human CD56(bright)  
214 natural killer cells contributes to renal fibrosis and chronic kidney disease progression.  
215 *Kidney Int* **92**, 79-88 (2017).
- 216 15. Nikolic-Paterson, D.J. CD4+ T cells: a potential player in renal fibrosis. *Kidney Int* **78**,  
217 333-335 (2010).
- 218 16. Crouse, J., Xu, H.C., Lang, P.A. & Oxenius, A. NK cells regulating T cell responses:  
219 mechanisms and outcome. *Trends Immunol* **36**, 49-58 (2015).
- 220 17. Kang, H.M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic  
221 variation. *Nat Biotechnol* **36**, 89-94 (2018).
- 222 18. Henig, N. et al. Interferon-beta induces distinct gene expression response patterns in  
223 human monocytes versus T cells. *PLoS One* **8**, e62366 (2013).
- 224 19. van Boxel-Dezaire, A.H. et al. Major differences in the responses of primary human  
225 leukocyte subsets to IFN-beta. *J Immunol* **185**, 5888-5899 (2010).

226 **FIGURE LEGENDS**

227  
228 **Figure 1** (a) Overview of the DESC framework. DESC starts with parameter initialization in which  
229 a stacked autoencoder is used for pretraining and learning a low-dimensional representation of  
230 the input gene expression matrix. The resulting encoder is then added to the iterative clustering  
231 neural network to cluster cells iteratively. The final output of DESC includes cluster assignment,  
232 the corresponding probabilities for cluster assignment for each cell, and the low-dimensional  
233 representation of the data. (b) Analysis of the single-cell data generated from midbrain in  
234 Drosophila. DESC not only identified the three types of Kenyon cells, which are detectable by the  
235 Louvain's method, but also identified cholinergic, glutamatergic, and GABAergic neurons, which  
236 are harder-to-separate by the Louvain's method reported in the original paper.  
237

238 **Figure 2** (a) Clustering of the mouse retina bipolar cells by different methods. The cells are  
239 colored by replication IDs. Cells from six replicates were processed in two different batches  
240 (Bipolar1-Bipolar4 are replicates from batch1, and Bipolar5-6 are replicates from batch 2). (b) KL-  
241 divergence for measuring of batch mixing of different methods. (c) Batch effect mixing is improved  
242 over iterations in DESC. (d) KL-divergence decreases over iterations in DESC, indicating that  
243 batch effect removal is improved over iterations.  
244

245 **Figure 3** (a) DESC clustering of the human kidney data. Cell types were determined based on  
246 known marker genes. Endo\_AVR: Endothelial Ascending Vasa Recta; Endo\_DVR: Endothelial  
247 Descending Vasa Recta; CD-IC: Collecting Duct Intercalated Cell; NK: Natural Killer; PT:  
248 Proximal Tubule. (b) KL-divergence for measuring batch mixing of different methods for the  
249 human kidney data. (c) DESC clustering of the PBMC data. Cell types were based on  
250 assignment reported in the original paper. (d) Volcano plot of differential expression analysis  
251 between control and stimulus conditions for each cell type. Highlighted are genes with FDR  
252 adjusted p-value< $10^{-50}$ . CD14<sup>+</sup> monocytes has the most number of differentially expressed  
253 genes compared to the other cell types. (e) Comparison of memory usage and running time of  
254 each method for datasets with various numbers of cells, where the cells were randomly sampled  
255 from the 1.3 million mouse brain dataset.

256 **ONLINE METHODS**  
257

258 **The DESC algorithm.** Analysis of scRNA-seq data often involves clustering of cells into different  
259 clusters and selection of highly variable genes for cell clustering. As these are closely related, it  
260 is desirable to use a data driven approach to cluster cells and select genes simultaneously. This  
261 problem shares similarity with pattern recognition, in which clear gains have resulted from joint  
262 consideration of the classification and feature selection problems by deep learning. However, for  
263 scRNA-seq data, a challenge is that we cannot train deep neural network with labeled data as  
264 cell type labels are typically unknown. To solve this problem, we take inspiration from recent work  
265 on unsupervised deep embedding for clustering analysis<sup>20</sup>, in which we iteratively refine clusters  
266 with an auxiliary target distribution derived from the current soft cluster assignment. This process  
267 gradually improves clustering as well as feature representation.  
268

269 *Overview of DESC.* The DESC procedure starts with parameter initialization, in which a stacked  
270 autoencoder is used for pretraining and learning low-dimensional representation of the input gene  
271 expression matrix. The corresponding encoder is then added to the iterative clustering neural  
272 network. The cluster centers are initialized by the Louvain's clustering algorithm<sup>5</sup>, which aims to  
273 optimize modularity for community detection. This clustering returns data in a feature space that  
274 allows us to obtain centroids in the initial stage of the iterative clustering. Below, we describe each  
275 component of the DESC procedure in detail.  
276

277 *Parameter initialization by stacked autoencoder.* Let  $X \in R^{n \times p}$  be the gene expression matrix  
278 obtained from a scRNA-seq experiment, in which rows correspond to cells and columns  
279 correspond to genes. Due to sparsity and high-dimensionality of scRNA-seq data, to perform  
280 clustering, it is necessary to transform the data from high dimensional space  $R^p$  to a lower  
281 dimensional space  $R^d$  in which  $d \ll p$ . Traditional dimension-reduction techniques such as  
282 principal component analysis, operate on a shallow linear embedded space, and thus have limited  
283 ability to represent the data. To better represent the data, we perform feature transformation by a  
284 stacked autoencoder, which have been shown to produce well-separated representations on real  
285 datasets.  
286

287 The stacked autoencoder network is initialized layer by layer with each layer being an  
288 autoencoder trained to reconstruct the previous layer's output. After greedy layer-wise training,  
289 all encoder layers are concatenated, followed by all decoder layers, in reverse layer-wise training  
290 order. The resulting autoencoder is then fine-tuned to minimize reconstruction loss. The final  
291 result is a multilayer autoencoder with a bottleneck layer in the middle. After fine tuning, the  
292 decoder layers are discarded, and the encoder layers are used as the initial mapping between  
293 the original data space and the dimension-reduced feature space, as shown in **Fig. 1a**.  
294

295 Since the number of true clusters for a scRNA-seq dataset is typically unknown, we apply the  
296 Louvain's method, a graph-based method that has been shown to excel over other clustering  
297 methods, on the feature space  $Z$  obtained from the bottleneck layer. This analysis returns the  
298 number of clusters, denoted by  $K$ , and the corresponding cluster centroids  $\{\mu_j : j = 1, \dots, K\}$ ,  
299 which will be used as the initial clustering for DESC.  
300

301 *Iterative clustering.* After cluster initialization, we improve the clustering using an unsupervised  
302 algorithm that alternates between two steps until convergence. In the first step, we compute a soft  
303 assignment of each cell between the embedded points and the cluster centroids. Following van  
304 der Maaten & Hinton<sup>21</sup>, we use the Student's  $t$ -distribution as a kernel to measure the similarity  
305 between embedded point  $z_i$  for cell  $i$  and centroid  $\mu_j$  for cluster  $j$ ,

306

307

$$q_{ij} = \frac{\left(1 + \|z_i - \mu_j\|^2 / \alpha\right)^{-1}}{\sum_{j'} \left(1 + \|z_i - \mu_{j'}\|^2 / \alpha\right)^{-1}}$$

308

309 where  $z_i = f_W(x_i) \in Z$  corresponds to  $x_i \in X$  after embedding,  $\alpha$  is the degree of freedom of the  
310 Student's  $t$ -distribution.

311

312 In the second step, we refine the clusters by learning from cells with high confidence cluster  
313 assignments with the help of an auxiliary target distribution. Specifically, we define the objective  
314 function as a Kullback-Leibler (KL) divergence loss between the soft cell assignments  $q_i$  and the  
315 auxiliary distribution  $p_i$  for cell  $i$  as

316

317

$$L = KL(P \parallel Q) = \sum_{i=1}^n \sum_{j=1}^K p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

318

319 where the auxiliary distribution  $P$  is defined as

320

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^n q_{ij}}{\sum_{j=1}^K (q_{ij}^2 / \sum_{i=1}^n q_{ij})}$$

321

322 The encoder is fine-tuned by minimizing  $L$ . The above definition of the auxiliary distribution  $P$  can  
323 improve cluster purity by putting more emphasis on cells assigned with high confidence. Given  
324 that the target distribution  $P$  is defined by  $Q$ , minimizing  $L$  implies a form of self-training. Also,  $p_{ij}$   
325 gives the probability of cell  $i$  that belongs to cluster  $j$ , and this probability can be used to measure  
326 the confidence of cluster assignment for each cell. Because  $\alpha$  is insensitive to the clustering result,  
327 we let  $\alpha = 1$  for all datasets by default.

328

329 *Optimization of the KL divergence loss.* We jointly optimize the cluster centers  $\{\mu_j: j = 1, \dots, K\}$   
330 and the deep neural network parameters using stochastic gradient descent. The gradients of  $L$   
331 with respect to feature space embedding of each data point  $z_i$  and each cluster center  $\mu_j$  are

332

$$\frac{\partial L}{\partial z_i} = \frac{\alpha + 1}{\alpha} \sum_{j=1}^K \left(1 + \frac{\|z_i - \mu_j\|^2}{\alpha}\right)^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j)$$

333

$$\frac{\partial L}{\partial \mu_j} = \frac{-(\alpha + 1)}{\alpha} \sum_{i=1}^n \left(1 + \frac{\|z_i - \mu_j\|^2}{\alpha}\right)^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j)$$

334

335 These gradients are then passed down to the deep neural network and used in standard  
336 backpropagation to compute the deep neural network's parameter gradient. We use Keras to train  
337 our model. During each iteration i.e. when loss is not decreasing or the epoch number threshold  
338 is reached, we update the auxiliary distribution  $P$ , and optimize cluster centers and encoder  
339 parameters with the new  $P$ . This iterative procedure is stopped when less than  $tol\%$  of cells  
340 change cluster assignment between two consecutive iterations. We let  $tol = 0.5$  by default.  
341

342 *Architecture of the deep neural network in DESC.* Depending on the number of cells in the dataset,  
343 we suggest different numbers of hidden layers and different numbers of nodes in the encoder.  
344 **Supplementary Table 2** gives the default numbers of hidden layers and nodes in DESC.

345  
346 DESC allows users to specify their own numbers of hidden layers and nodes. We recommend  
347 using more hidden layers and more nodes per layer for datasets with more cells so that the  
348 complexity of the data can be captured by the deep neural network. We use ReLU as the  
349 activation function except for the last hidden layer and last decoder layer, in which tanh is used  
350 as the activation function. The reason why we use tanh is that we must guarantee the values in  
351 feature representation and output of decoder range from negative to positive. The default  
352 hyperparameters for the autoencoder are listed in **Supplementary Table 3**.

353  
354 **Data normalization and gene selection.** The normalization involves two steps. In the first step,  
355 cell level normalization is performed, in which the UMI count for each gene in each cell is divided  
356 by the total number of UMIs in the cell, and then transformed to a natural log scale. In the second  
357 step, gene level normalization is performed in which the cell level normalized values for each  
358 gene are standardized by subtracting the mean across all cells and divided by the standard  
359 deviation across all cells for the given gene. Highly variable genes are selected using the  
360 filter\_genes\_dispersion function from the Scanpy package<sup>22</sup> (<https://github.com/theislab/scanpy>).  
361

362 **Evaluation metric for clustering.** For published datasets in which the reference cell type labels  
363 are known, we use ARI to compare the performance of different clustering algorithms. Larger  
364 values of ARI indicate higher accuracy in clustering. The ARI can be used to calculate similarity  
365 between the clustering labels obtained from a clustering algorithm and the reference cluster labels.  
366 Given a set of  $n$  cells and two sets of clustering labels of these cells, the overlap between the two  
367 sets of clustering labels can be summarized in a contingency table, in which each entry denotes  
368 the number of cells in common between the two sets of clustering labels. Specifically, the ARI is  
369 calculated as  
370

$$ARI = \frac{\sum_{jj'} \binom{n_{jj'}}{2} - [\sum_j \binom{a_j}{2} \sum_{j'} \binom{b_{j'}}{2}] / \binom{n_{jj'}}{2}}{\frac{1}{2} [\sum_j \binom{a_j}{2} + \sum_{j'} \binom{b_{j'}}{2}] - [\sum_j \binom{a_j}{2} \sum_{j'} \binom{b_{j'}}{2}] / \binom{n_{jj'}}{2}}$$

371  
372 where  $n_{jj'}$  is the number of cells assigned to cluster  $j$  based on the reference cluster labels, and  
373 cluster  $j'$  based on clustering labels obtained from a clustering algorithm,  $a_j$  is the number of cells  
374 assigned to cluster  $j$  in the reference set, and  $b_{j'}$  is the number of cells assigned to cluster  $j'$  by  
375 the clustering algorithm.  
376

377  
378 **Evaluation metric for batch effect removal.** We use KL-divergence to evaluate the performance  
379 of various single-cell clustering algorithms for batch effect removal i.e., how randomly are cells  
380 from different batches mixed together within each cluster. The KL-divergence of batch mixing for  
381  $B$  different batches is calculated as  
382

$$KL = \sum_{b=1}^B p_b \log \frac{p_b}{q_b}$$

383  
384

385 where  $q_b$  is the proportion of cells from batch  $b$  among all cells, and  $p_b$  is the proportion of cells  
386 from batch  $b$  in a given region based on results from a clustering algorithm, with  $\sum_{b=1}^B q_b = 1$  and  
387  $\sum_{b=1}^B p_b = 1$ . We calculate the KL divergence of batch mixing on the first two components of the  
388 t-SNE coordinates, by using regional mixing KL divergence defined above at the location of 100  
389 randomly chosen cells from all batches. The regional proportion of cells from each batch is  
390 calculated based on the set of 120 nearest neighboring cells from each randomly chosen cell.  
391 The final KL divergence is then calculated as the average of the regional KL divergence. We  
392 repeated this procedure for 500 iterations with different randomly chosen cells to generate box  
393 plots of the final KL divergence. Smaller final KL divergence indicates better batch mixing i.e.,  
394 more effective in batch effect removal.  
395

396 **Datasets.** We analyzed multiple scRNA-seq datasets. Publicly available data were acquired from  
397 the access numbers provided by the original publications. The human kidney dataset generated  
398 by us is available in Supplementary Data.  
399

400 *Benchmarking dataset.* The Genotype-Tissue Expression (GTEx) v7 dataset<sup>10</sup> was downloaded  
401 from the GTEx data portal (<https://gtexportal.org/home/datasets>). This dataset includes 11,688  
402 human RNA-seq samples from 30 tissues. Because the tissue origin is known, we treat this  
403 dataset as the benchmarking dataset in which the tissue origin is used as the true cluster label.  
404

405 *Drosophila dataset.* The data were generated by Croset et al.<sup>11</sup> in which 10,286 cells were  
406 generated using Drop-seq from the midbrain of drosophila.  
407

408 *Mouse retina dataset.* The data were generated by Shekhar et al.<sup>12</sup> in which 23,494 bipolar cells  
409 were generated using Drop-seq from retinas of six mice processed in two experimental batches.  
410 This dataset allows us to examine batch effect at the subject level.  
411

412 *Human kidney datasets.* The first set of data was generated by us using 10X. This dataset  
413 includes 8,544 cells from kidneys in four healthy human subjects. The second set of data was  
414 generated by Young et al.<sup>13</sup>, also using 10X. This dataset includes 7,149 cells from the normal  
415 part of the kidneys in three human subjects that have kidney tumors. These two datasets were  
416 combined in our analysis, which allow us to examine batch effect, both at the subject level and at  
417 the dataset level.  
418

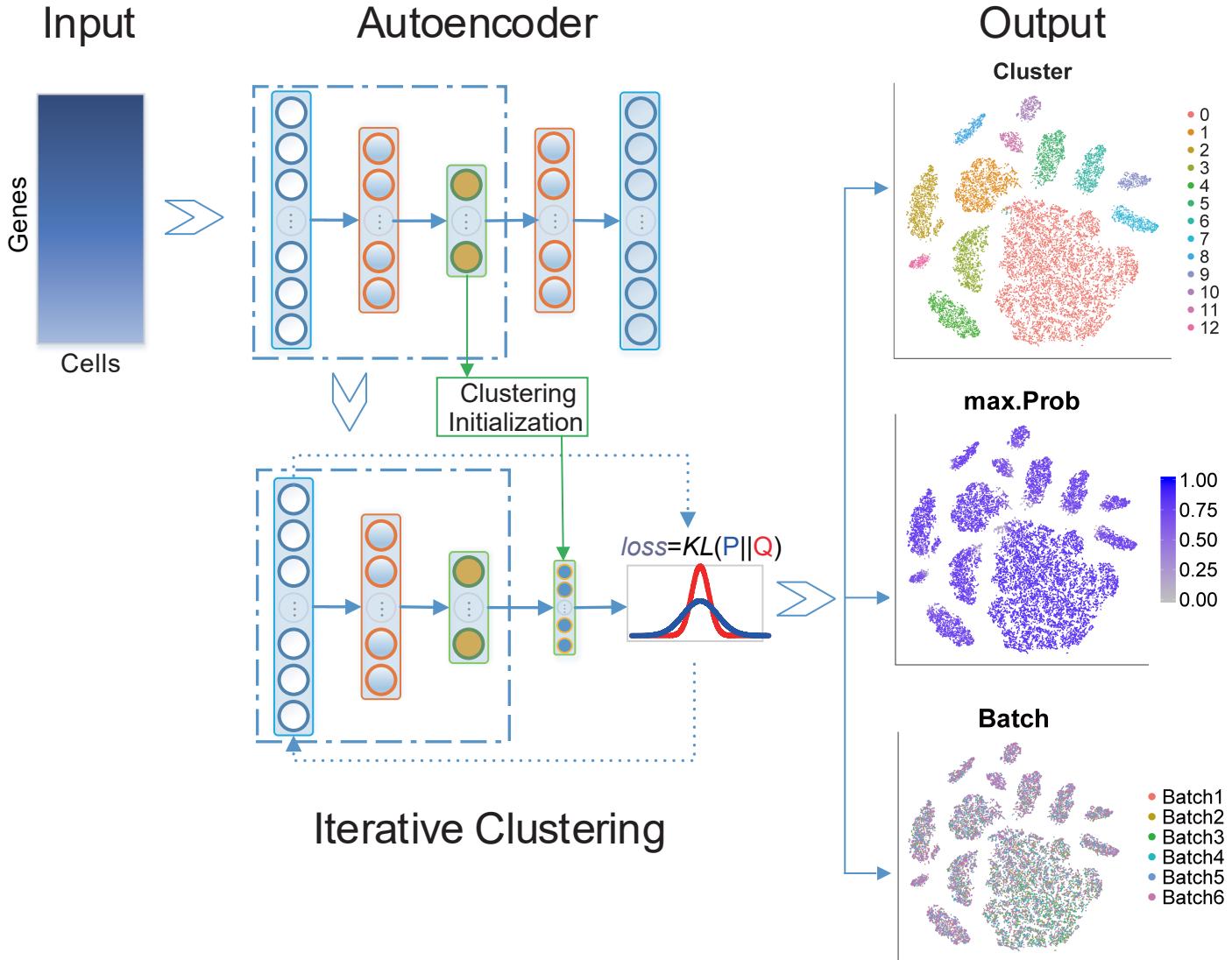
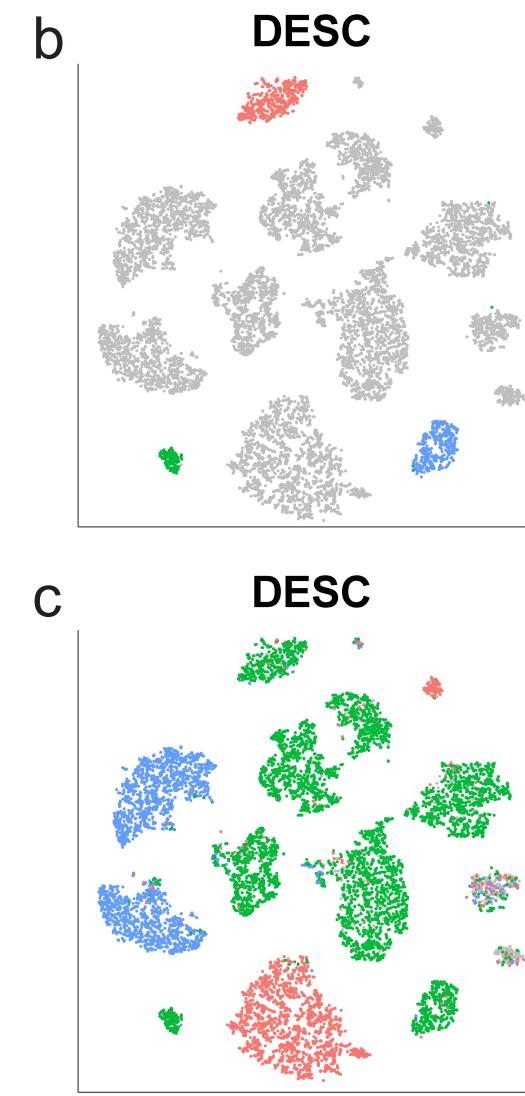
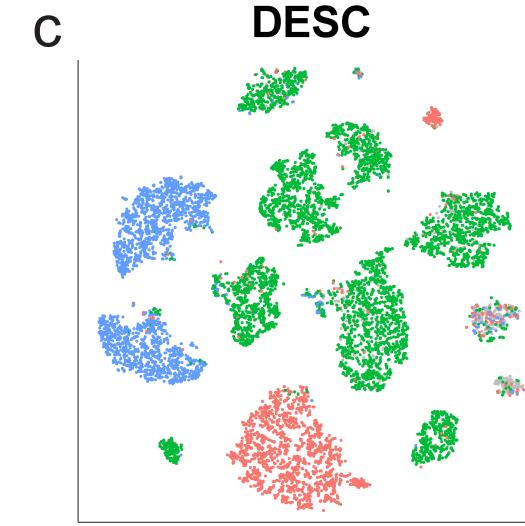
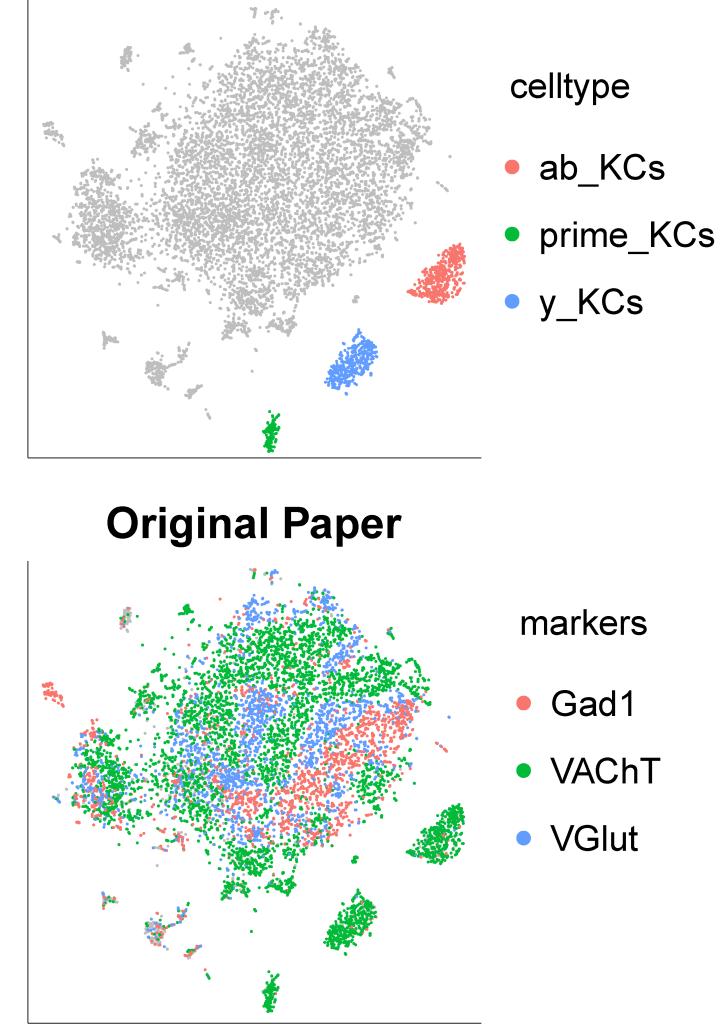
419 *Human PBMC dataset.* The data were generated by Kang et al.<sup>17</sup> in which 24,679 PBMC cells  
420 were obtained and processed from eight patients with lupus using 10X. These cells were split into  
421 two groups: one stimulated with interferon-beta (INF-β) and a culture-matched control. This  
422 dataset allows us to examine whether technical batch effect can be removed in the presence of  
423 true biological variations.  
424

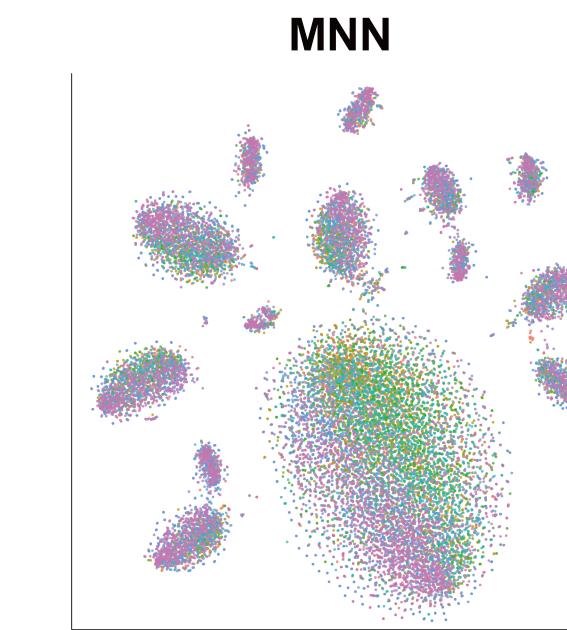
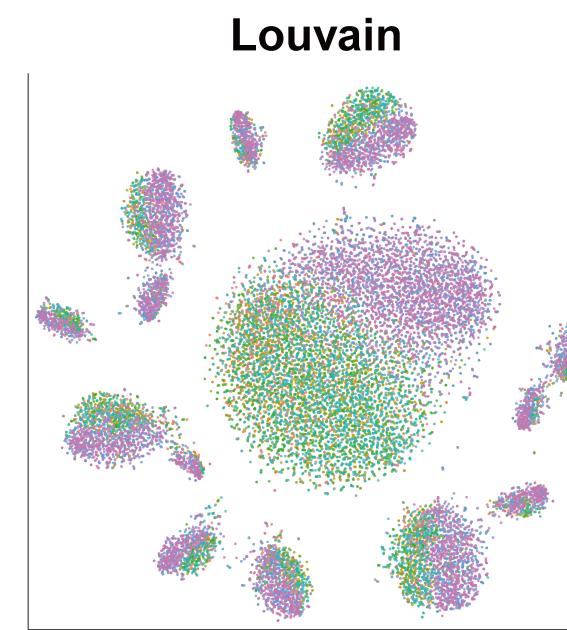
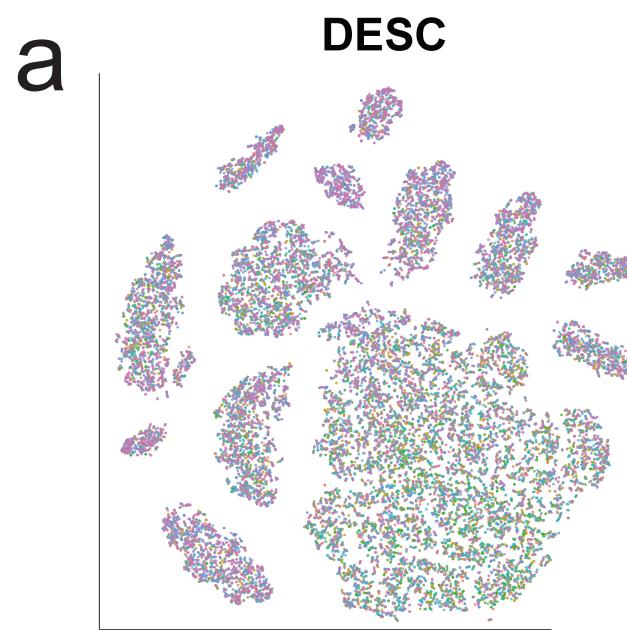
425 *1.3 million brain cells from E18 mice.* This dataset was downloaded from the 10X Genomics  
426 website ([https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M\\_neurons](https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons)). It includes 1,306,127 cells from cortex, hippocampus and subventricular zone of two  
427 E18 C57BL/6 mice.  
428

429 A complete list of the datasets analyzed in this paper is provided in **Supplementary Table 1**.  
430

431 **Software availability.** An open-source implementation of the DESC algorithm can be  
432 downloaded from <https://eleozzr.github.io/desc/>.  
433

- 435 20. Xie, J., Girshick, R. and Farhadi, A. Unsupervised deep embedding for clustering analysis.  
436 Proceedings of the 33<sup>rd</sup> International Conference on Machine Learning *JMLR: W&CP* **48** (2016).  
437  
438 21. van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *JMLR* **9**, 2579-2605 (2008).  
439  
440 22. Wolf, F.A., Angerer, P. and Theis, F.J. SCANPY: large-scale single-cell gene expression  
441 data analysis. *Genome Biology* **19**, 15 (2018).

**a****b****c****Original Paper**



**BatchID**

- Bipolar1
- Bipolar2
- Bipolar3
- Bipolar4
- Bipolar5
- Bipolar6

