Supplementary

PROJECT CELLS AND GENES IN THE SAME SPACE

For the Guo data set, after we got the 200 dimensional representations for cells and genes simultaneously by SCRL, we used PCA to project the cells in 2D and 3D space respectively. Because the genes and cells shared the same 200 dimensional space, we performed the same line transformation for several key genes: POU5F1, NANOG, KIT, ALPL, SOX17 and CD38. Among them, POU5F1 and NANOG are pluripotency marker genes, which are expressed both in ICM cells and mitotic PGC cells. KIT, ALPL, SOX17 and CD38 are PGC specific marker genes [13]. As shown in **Figure S2** and **Figure S3**, these genes are colored by yellow, we can observe that these genes are more closely linked to PGC cells, which is as expected. In order to measure the distance between the cells and the genes more quantitatively, we calculated the Euclidean distance between each gene and each cell type's cluster center in 3D. Each cell type's cluster center is defined as the mean representation of this cell type in 3D. The results are shown in **Table S1**. The quantification results are consistent with the intuitive visualization.

In addition, the expression pattern of these six marker genes were shown in **Figure S4.**

TIME COMPLEXITY ANALISIS OF SCRL

By adopting the alias method [1], sampling an edge whose weight is greater than 0 from a bipartite network only takes O(1) time. We found that the number of iterations T should be proportional to $X = \max(|E_{ca}|,|E_{ga}|)$ in practice.(Here, we found T = 4X is enough for all the three datasets in this paper. So we use T = 4X for the following runtime comparison analysis.)

Therefore, the overall time complexity of our SCRL model is O(LKX). Where K is the number of negative samples, L is the dimension of the low dimensional representations, $|E_{ca}|$ is the number of edges in Cell-ContextGene network, $|E_{ga}|$ is the number of edges in Gene-ContextGene network. So the edge-sampling based optimization method ensures the efficiency and effectiveness, which is able to handle with large data sets. In addition, SCRL can be executed with multi threads. We recorded the runtime for the four dimensional reduction methods on these three data sets. The average runtime is used based on five repeated runs. The results are shown in **Figure S5**. SCRL has comparable runtimes with PCA and it runs much faster than t-SNE & ZIFA in the three studied datasets. All the models compared in this paper were run on a single machine with 128G memory and 16 CPU cores.

We also investigated the relationship between the runtime and the thread number. Results in **Figure S6** show that the runtime first decreases and finally becomes stable when the thread number increases from 1 to 10. So we set the thread number to be 10 for all the three datasets in our paper.

REFERENCES

1. Walker, A. J. (1974). New fast method for generating discrete random numbers with arbitrary frequency distributions. Electronics Letters, 10(8), 127-128.

FIGURES AND TABLES

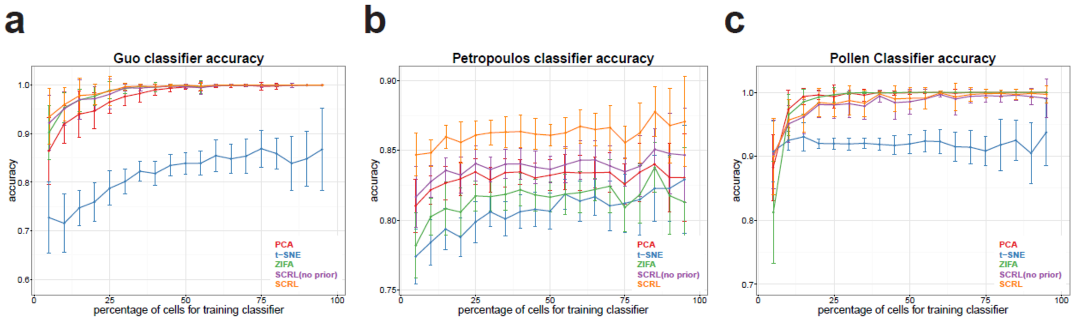

Figure S1 – Performance comparison of the 4 dimensional reduction methods for Classification accuracy on the three data sets. (a) Guo data set. (b) Petropoulos data set. (c) Pollen data set. The x axis corresponds to the percentage (from 5% to 95%) of the cells for training classifier, each color represents one method. The y axis represents the classification accuracy for the test data.

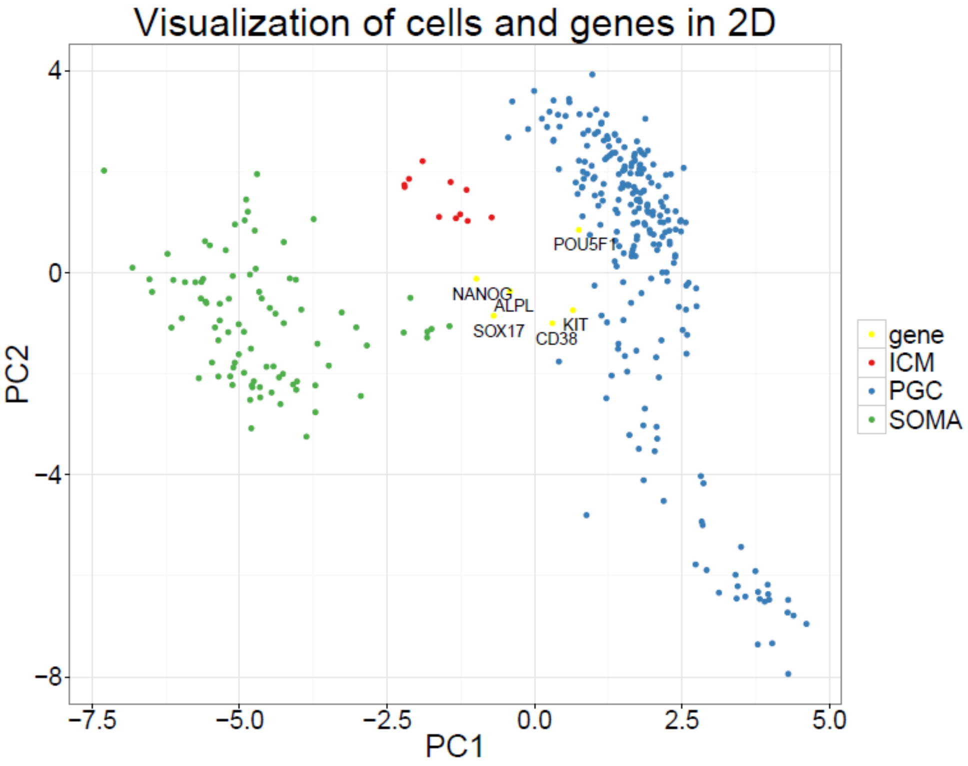

Figure S2 –Visualization of cells and genes in 2D for Guo's data set. The red, the blue and the green points represent the cells, the yellow point represents a gene. Each single cell is colored according to its true cell type label.

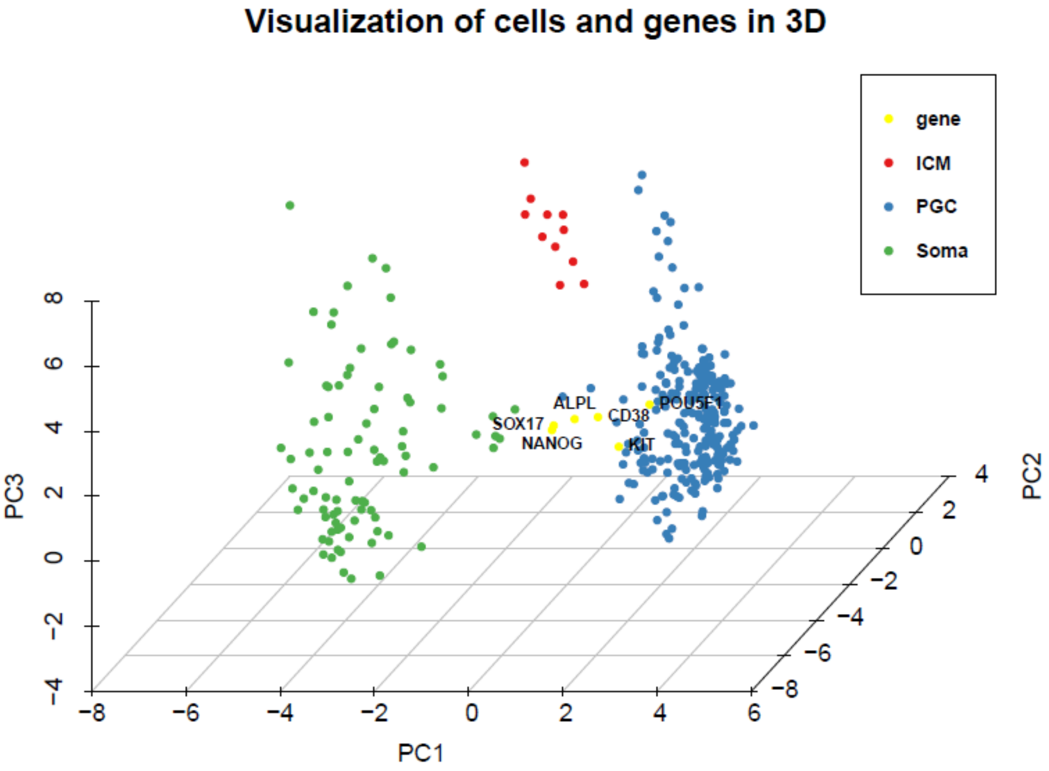

Figure S3 –Visualization of cells and genes in 3D for Guo's data set. The red, the blue and the green points represent the cells, the yellow point represents a gene. Each single cell is colored according to its true cell type label.

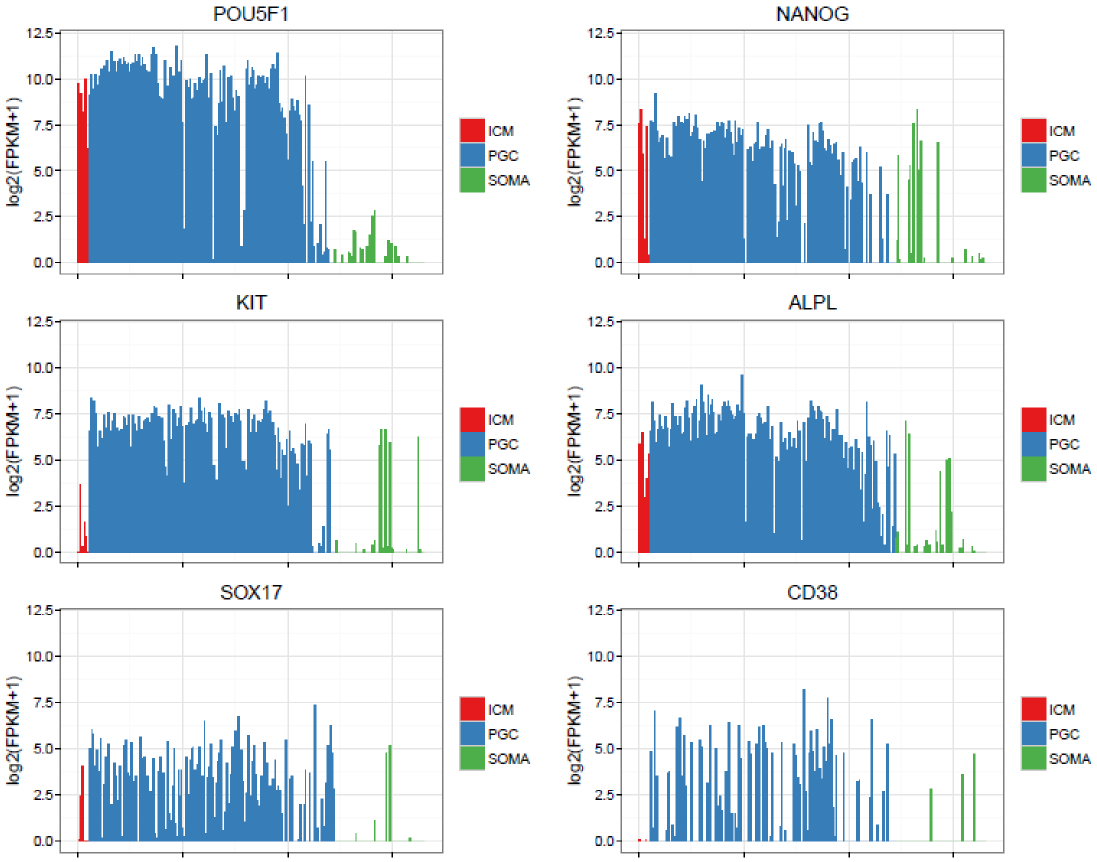

Figure S4 – Expression patterns of selected marker genes for Guo's data set. The x axis represents the cell, the y axis represents the expression value log2(FPKM+1) of different marker genes. Each single cell is colored according to its true cell type label.

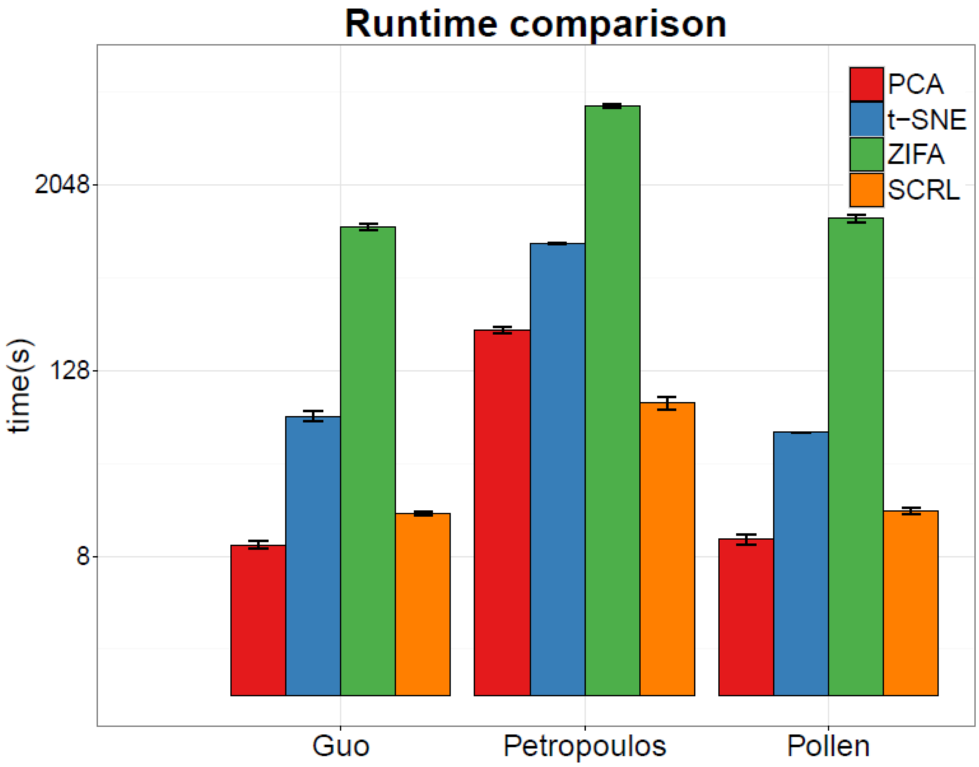

Figure S5 – Runtime comparison of SCRL, PCA, t-SNE and ZIFA. Each color corresponds to one method, the x axis represents different data sets, the y axis represents the time(s). s represents the second.

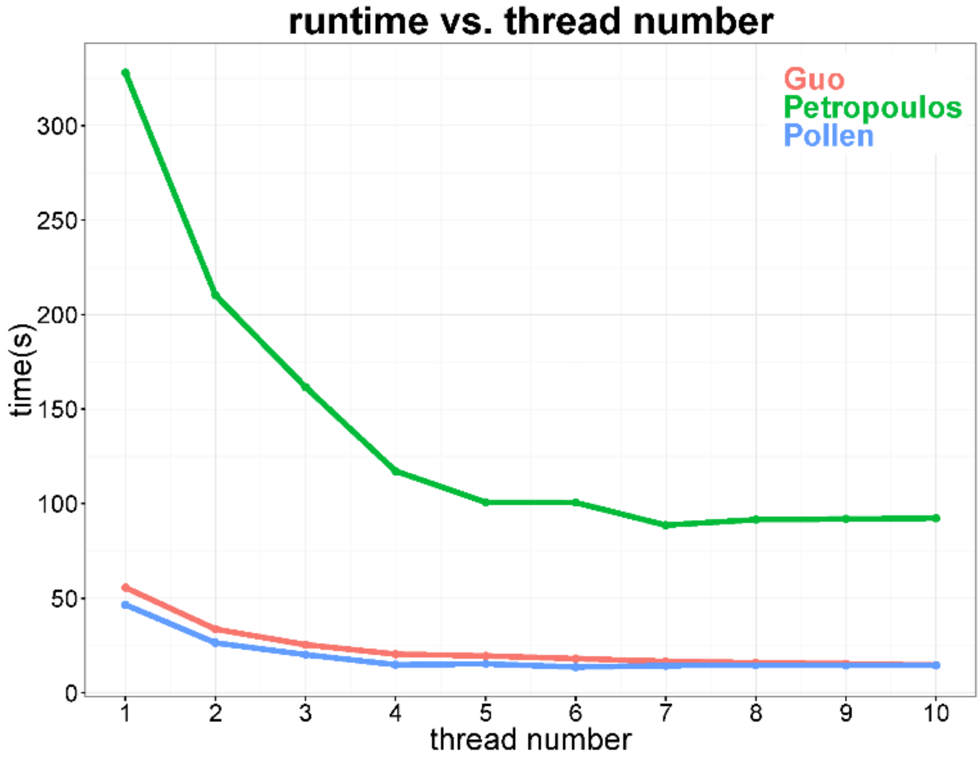

Figure S6 – Runtime v.s. thread number. Each color corresponds to one data set, the x axis represents the thread number, the y axis represents the runtime.

	ICM	PGC	SOMA
POU5F1	5.499689	1.185890	5.722889
NANOG	5.478388	2.819485	3.775041
KIT	6.222548	1.587096	5.320110
ALPL	5.208420	2.316836	4.331407
SOX17	5.288097	2.748347	4.032784
CD38	5.319823	2.073151	5.069109

Table S1 – Euclidean distance between selected marker genes and cell types in 3D.

Each column corresponds to one cell type, each row corresponds to one selected marker gene. The value is the Euclidean distance between one specific cell type and one selected marker gene in 3D.